SEMICONDUCTOR DEVICES

V. Stupelman
and G. Filaretov

The book discusses the principles of most widely used semiconductor devices and physical processes determining their characteristics. The coverage also includes the design, electric parameters and fields of application of semiconductors. Much consideration is given to the design features and principles of fabrication of a variety of integrated circuits. Detailed analysis is presented on the modern state of semiconductor electronics and the trends of its development. For a better understanding of semiconductor operation, the first two chapters are devoted to the contact phenomena in *p-n* junctions. The book is intended for students of technical schools and will also be useful to college students.

ИЗДАТЕЛЬСТВО • МИР •

ПОЛУПРОВОДНИКОВЫЕ ПРИБОРЫ

В. Ш. Ступельман
Г. А. Филаретов

Издательство «Советское радио»
Москва

SEMICONDUCTOR DEVICES

V. Stupelman
and G. Filaretov

Translated
from the Russian
by
P. Ivanov

Mir Publishers
Moscow

First published 1976
Revised from the 1973 Russian edition

The Russian Alphabet and Transliteration

Аа	a	Кк	k	Хх	kh
Бб	b	Лл	l	Цц	ts
Вв	v	Мм	m	Чч	ch
Гг	g	Нн	n	Шш	sh
Дд	d	Оо	o	Щщ	shch
Ее	e	Пп	p	ъ	″
Ёё	ё	Рр	r	ы	y
Жж	zh	Сс	s	ь	′
Зз	z	Тт	t	Ээ	e
Ии	i	Уу	u	Юю	yu
Йй	y	Фф	f	Яя	ya

The Greek Alphabet

Αα	Alpha	Ιι	Iota	Ρρ	Rho
Ββ	Beta	Κκ	Kappa	Σσ	Sigma
Γγ	Gamma	Λλ	Lambda	Ττ	Tau
Δδ	Delta	Μμ	Mu	Υυ	Upsilon
Εε	Epsilon	Νν	Nu	Φφ	Phi
Ζζ	Zeta	Ξξ	Xi	Χχ	Chi
Ηη	Eta	Οο	Omicron	Ψψ	Psi
Θϑθ	Theta	Ππ	Pi	Ωω	Omega

На английском языке

© English translation, Mir Publishers, 1976

CONTENTS

Preface 9

List of Symbols 10

Introduction 15

Chapter One. **Contact Phenomena in Semiconductors** 19

1.1. Contact Potential Difference 19
1.2. Current Rectification at Semiconductor-Metal Contact 23
1.3. Rectification at Metal-Semiconductor Contact 27

Review Questions 32

Chapter Two. **Electron-Hole Junction** 33

2.1. Types of *p-n* Junctions. Physical Model of a *p-n* Junction 33
2.2. Calculation of the Width of Space-Charge Layer in the *p-n* Junction 37
2.3. Rectification at the *p-n* Junction 40
2.4. Calculation of the Current-Voltage Characteristic of a *p-n* Junction 43
2.5. Effect of the Injection Level and Temperature on the Current-Voltage Characteristic of a *p-n* Junction 46
2.6. Field Strength and Breakdown of the *p-n* Junction 50
2.7. *p-n* Junction Capacitance 54
2.8. Ohmic Contact 56
2.9. Surface State Properties of a Semiconductor 57

Review Questions 59

Chapter Three. **Diodes** 61

3.1. Diode Applications and Classification 61
3.2. Methods of Producing *p-n* Junctions in Diodes 62
3.3. Equivalent Circuit and Parameters of Diodes 66

3.4. Germanium and Silicon Rectifier Diodes . . 70
3.5. Polycrystalline Rectifiers 75

Review Questions 77

Chapter Four. **Special Diodes** 78

4.1. Switching Diodes 78
4.2. Zener Diodes 82
4.3. Varicaps 87
4.4. Tunnel Diodes 92
4.5. Gunn and Impatt Diodes 97

Review Questions 102

Chapter Five. **Transistor: Principles and Characteristics** 103

5.1. Transistor as an Amplifier 103
5.2. Calculation of Current Gain 105
5.3. Intrinsic Resistances of the Transistor . . . 109
5.4. Transistor Currents and Voltages 114
5.5. Transistor Capacitance 116
5.6. Frequency Dependence of the Current Gain 117
5.7. Transistor Noise 119

Review Questions 120

Chapter Six. **Transistor as a Fourpole** 122

6.1. Transistor as an Active Fourpole 122
6.2. Three Basic Transistor Configurations 126
6.3. Static Characteristics of the Transistor 131
6.4. Dependence of Transistor Main Parameters on Operating Conditions and Temperature . . . 133
6.5. Pulse (Switching) Response of the Transistor 136
6.6. Maximum Permissible Operating Conditions of the Transistor 141

Review Questions 143

Chapter Seven. **Transistor Types and Applications** 144

7.1. Classification of Transistors 144
7.2. Alloy Transistors 147
7.3. Drift Transistors 153
7.4. Calculation of the Planar Transistor 162

Review Questions 167

Chapter Eight. **Field-Effect and Unijunction Transistors. *p-n-p-n* Switches** 168
8.1. Field-Effect Transistor 168
8.2. Insulated-Gate Field-Effect Transistor 171
8.3. Unijunction Transistor 179
8.4. *p-n-p-n* Switch 180

Review Questions 185

Chapter Nine. **Thermoelectric and Galvanomagnetic Devices** 186
9.1. Thermoelectric Phenomena in Semiconductors 186
9.2. Thermoelectric Generators 190
9.3. Thermoelectric Coolers 194
9.4. Hall-Effect Devices 197

Review Questions 202

Chapter Ten. **Photodevices** 203
10.1. Photoconduction 203
10.2. Photoresistors 210
10.3. Photodiodes and Photocells 213
10.4. Phototransistors 219
10.5. Semiconductor Light Sources 221

Review Questions 226

Chapter Eleven. **Integrated Circuits** 227
11.1. Reliability of Semiconductor Devices . . . 227
11.2. Major Objectives of Microelectronics 230
11.3. Main Principles of Preparation of Integrated Circuits 234
11.4. Semiconductor Integrated Circuits 236
11.5. Film Integrated Circuits 238
11.6. Epitaxial-Film Integrated Circuits 245

Review Questions 247

Chapter Twelve. **Semiconductor Technology Today. Trends of Development** 248

Appendix 1. Fundamental Constants 256

Appendix 2. Properties of Semiconductors at 300 °K 256

Appendix 3. Thermo-emf of Semiconductors 257
Appendix 4. Thermo-emf of Metals and Alloys 257
Appendix 5. Relationships among SI and cgs Units 258
Appendix 6. Standard Graphical Symbols for Semiconductor Devices 258

References 262
Index 263

PREFACE

The book has been written in accordance with the course "Semiconductor devices" for electronic technical schools. It may also be found useful for students of technical colleges and engineers engaged in the production and application of semiconductor devices.

The volume is based on the lectures read by the authors at the Kalinin Polytechnical Institute of Leningrad.

The book deals with the principles of operation, design and major characteristics of modern semiconductor devices. Most popular semiconductor devices, such as diodes, bipolar and field-effect transistors are treated in greater detail. Much attention is given to the problems of microelectronics because of its promising advantages and exceptional importance. The limited space of the book has not allowed the authors to consider varistors, thermistors and strain gauges.

The first two chapters are devoted to the physics of contact phenomena and the properties of *p-n* junctions, which are indispensable for understanding the physical processes occurring in semiconductor devices.

Bearing in mind the inextricable relation between design and technology, which is particularly important in the manufacture of semiconductor devices, the authors have found it necessary to touch upon a number of technological problems. The book does not consider the aspect of measurement of semiconductor parameters since the subject is treated in sufficient detail in other college courses.

The book has been written on the initiative of N. A. Novoselov, director of the Leningrad radio engineering and polytechnical school, and R. A. Gavrilov.

The authors wish to thank the reviewers of the book E. I. Zhdanova and Yu. A. Ovechkin, Cand. Sc. (Tech.), and also the scientific editor A. N. Karmazinsky, associate professor, Cand. Sc. (Tech.), for their valuable suggestions.

LIST OF SYMBOLS

B magnetic induction
b ratio of electron mobility to hole mobility
C capacitance
C_{bar} barrier (charge) capacitance
C_{dif} diffusion capacitance
C_g gate capacitance
c heat capacity
D diffusion constant for impurities
D_n diffusion constant for electrons
D_p diffusion constant for holes
d width of channel or width of semiconductor sample
E electric field strength; emf of supply source
E_F Fermi energy
E_{Fs} Fermi energy at surface
E_i energy in the middle of band gap (forbidden band)
ΔE band gap width
E_n energy of electron
E_p energy of hole
f frequency
f_α (f_{h21b}) cutoff frequency in common-base connection
f_β (f_{h21e}) cutoff frequency in common-emitter connection
H magnetic field strength
h Planck's constant
I current
I_e, I_c, I_b emitter, collector and base current, respectively
I_g, I_s, I_d gate, source and drain current, respectively
I_0 mean value of rectified current; thermal current
I_{for} forward current of diode
I_{rev} reverse current of diode
I_{c0} reverse collector current with base-emitter circuit open
I_p peak current of tunnel diode
I_v valley current of tunnel diode
I_{st} regulation (stabilization) current
I_{ph} photocurrent
J current density

J_{rec} recombination current density
k Boltzmann's constant; light absorption coefficient
L_n diffusion length for electrons
L_p diffusion length for holes
L incident radiation intensity
l length of channel
M multiplication factor for charge carriers
m_0 mass of free electron
m^* effective mass of electron
N_a acceptor density
N_d donor density
N_c density of states in the conduction band
N'_a density gradient of acceptors
N'_d density gradient of donors
n electron density
n_b density of electrons in semiconductor bulk
n_s density of electrons at semiconductor surface
n_p equilibrium density of electrons in p-region
n_i intrinsic density of charge carriers in semiconductor
P dissipated power; Peltier coefficient
p hole density
p_b hole density in semiconductor bulk
p_s hole density at semiconductor surface
p_n equilibrium density of holes in n-region
Q heat quantity; quality factor of voltage variable capacitor
Q_z quality factor of Zener (avalanche) diode
q electronic charge
q_n cross section of electron capture
q_p cross section of hole capture
R Hall constant; electric resistance
R_{th} thermal resistance of (Hall) device
R_d differential resistance of (Hall) device
R_l load resistance
r_e, r_b, r_c emitter, base and collector resistance, respectively
r_r resistance of semiconductor bulk
S cross-sectional area, transadmittance
s surface recombination velocity; saturation coefficient
T absolute temperature
TCV temperature coefficient of voltage
TCR temperature coefficient of resistance
TCC temperature coefficient of capacitance
t time
t_{on} turn-on time
t_{off} turn-off time

t_{for} transient time for forward resistance
t_{rev} recovery time for reverse resistance
t_j temperature of junction
V external voltage
V_e, V_b, V_c emitter, base and collector voltage, respectively
V_g, V_s, V_d, V_0 gate, source, drain and threshold voltage, respectively
V_z, V_a, V_{th} Zener (tunnel), avalanche and thermal breakdown voltage, respectively
V_{st} regulation (stabilization) voltage
V_{for} forward voltage drop across diode
V_{rev} reverse voltage
V_p peak voltage of tunnel diode, punch-through voltage
V_v valley voltage of tunnel diode
v velocity of charge carrier
W base width, thickness of channel in unipolar transistors
α (h_{21b}) common-base current gain
α_T thermoelectric coefficient
β (h_{21e}) current-transfer ratio (in common-emitter connection)
β quantum yield
β^* transport efficiency (transport factor) of minority carriers across the base
γ injection efficiency, recombination coefficient
γ_H sensitivity of Hall device
γ^* emitter (injection) efficiency
δ width of space charge layer
δ^* injection level
δ_n space-charge layer width in n-semiconductor
δ_p space-charge layer width in p-semiconductor
ε relative permittivity of material
ε_0 permittivity of empty space
η efficiency
$\varkappa$ thermal conductivity
λ free path length; wavelength; failure rate
μ mobility of carriers of current; permeability
ν frequency of electromagnetic oscillations, condensation rate
ρ resistivity, space-charge density
σ conductivity of semiconductor
τ lifetime of charge carriers; Thomson coefficient
Φ light flux
φ_n electrostatic potential in n-semiconductor
φ_p electrostatic potential in p-semiconductor
φ_F Fermi potential

φ_S work function of electron in semiconductor
φ_m work function of electron in metal
$\Delta\varphi_d$ potential barrier (height) of *p-n* junction in equilibrium state; contact-potential difference
φ_s potential at surface
φ_b potential in semiconductor bulk
φ_T thermal potential
$\Delta\psi$ phase shift
ω angular frequency

INTRODUCTION

Modern semiconductor electronics is considered to have originated from the discovery of the transistor effect by J. Bardeen and W. Brattain in 1948. This discovery and the ensuing rapid development of the solid-state technology are the result of the numerous works by the scientists of many countries.

As long ago as 1888 the Russian scientist V. A. Ulyanin investigated the nature of a photoelectromotive force in selenium. In a number of countries semiconductor crystal detectors (using silicon carbide, silicon, tellurium crystals) with pressure wire contacts were studied and put to use. In 1922 O. V. Losev realized the idea of amplification and generation of electric oscillations through the use of semiconductor detectors displaying a falling portion on the current-voltage characteristics. The German scientist I. E. Lilienfeld was granted a patent in 1925 for an amplifying device using a cuprous sulphide film as a basic material. This device became the prototype of the modern field-effect transistor. In 1926 Grondahl devised a cuprous oxide rectifier. In 1938 the German scientists R. Hilsch and R. Pohl built a current control device employing a gallium bromide crystal.

In the twenties and thirties of this century extensive investigations were conducted in the field of semiconductors in a great many of countries. In the USSR a large group of physicists headed by the academician A. F. Ioffe achieved an appreciable success in the study of semiconductor properties. In 1932 I. E. Tamm developed the theory of surface states; in 1931-1934 B. V. Kurchatov, V. P. Zhuze and others established the relationship between the conductivity type of various semiconductors and the impurity concentration. Ya. I. Frenkel created the theory of excitation of paired charges (electron-hole pairs) in semiconductors. In 1938 B. I. Davydov developed the theory of rectification and determined the role of minority carriers in a semiconductor with a barrier layer.

But despite these great advances, the development rate of semiconductor devices could hardly be compared with that of vacuum devices. It was not until the beginning of the forties that intensive work had started on the industrial application of semiconductors because of the growing demands of television and radar for more advanced devices. New methods were developed for the preparation of such semiconductor materials as germanium, silicon, A^{III} and

B^V compounds, and also methods of their purification, mechanical and chemical treatment. These were the achievements that paved the way for the development of a point-contact transistor by J. Bardeen and W. Brattain in 1948, a junction transistor by W. Shockley in 1949 and, above all, for their commercial production. Already in 1955 the advanced countries turned out a great number of transistors embracing a power range from 50 mW to tens of watts and a working frequency range from tens of kilohertz to tens of megahertz.

Commercial diffused-alloy transistors of small and medium power, elaborated in the USSR in 1957-1958, along with the earlier developed alloy transistors made it possible to devise new types of fully transistorized radio equipment.

In the fifties new classes of semiconductor devices were worked out, such as tunnel diodes, field-effect transistors, thyristors and many others. At the beginning of the sixties the planar-epitaxial technology came into practice, which enabled serial production of high-frequency diodes and transistors with a narrow spread of parameters, high stability and a broad range of operating powers up to hundreds of watts. This technology allowed the production of complete radio circuits with tens and hundreds of active and passive components grown on a single crystal, and thus predetermined great advances in the field of microminiaturization.

Optical *p-n* junction masers created by the Soviet scientists N. Basov and A. Prokhorov gave a new impetus to the advancement of light-emitting semiconductor devices.

The modern stage of development of solid-state technology is noted for a rapid introduction of scientific achievements into practice. Just a few years have passed since the invention of semiconductor lasers, but now they are used to cut various materials, weld miniature parts and perform most delicate surgical operations.

Apart from the fact that semiconductor devices have formed the basis of modern radioelectronics and found extensive application in the computing technique, electrical engineering and various industries directly associated with the conversion of electrical energy, they find their way into the spheres bearing the remotest relation to electronics, such as biology, hydraulic engineering and agriculture, all this owing to such merits as low power consumption, small dimensions, a wide range of control and controlled parameters, the ability of direct conversion of the energy of one state to another (for example, conversion of heat energy directly into electrical energy in thermoelectric generators).

Semiconductor electronics has formed the basis for the creation of modern control systems and radio devices. Both in the USSR and other countries in progress is the large-scale research, experimental, design and technological work directed towards further development of semiconductor electronics.

The Directives of the 25th Congress of the CPSU on the 1976-1980 five-year plan of the development of the USSR national economy defined a vast programme on the implementation of the main objective — ensuring still higher living and cultural standards of the people on the basis of the accelerated development of the socialist economy, increased efficiency, scientific and technical progress and growing labour productivity. In this connection semiconductor electronics acquires an ever growing importance in solving the stipulated tasks.

Chapter One

CONTACT PHENOMENA IN SEMICONDUCTORS

1.1. CONTACT POTENTIAL DIFFERENCE

Let us consider an extrinsic semiconductor containing donor and acceptor impurities. At room temperature the impurity atoms in most semiconductors are ionized, and the positive charge of donor atoms and negative charge of acceptor atoms are balanced by the charge of free electrons and that of holes, respectively.

The electron density in the conduction band and the hole density in the valence band are related to the position of the Fermi level.

The Fermi level E_F or the mean energy of electrons in a semiconductor determines the probability of occupying the energy states. Thus, the energy levels E below E_F are occupied while the energy levels that lie above E_F by the value kT are empty. The probability of occupying the Fermi level proper is equal to one half.

The electron and hole densities in a semiconductor are related to the Fermi level by

$$n = n_i e^{\frac{E_F - E_i}{kT}}, \quad p = n_i e^{\frac{E_i - E_F}{kT}}$$

where $E_i = (E_c - E_v)/2$ is the energy level in the middle of the band gap (forbidden band) of the semiconductor or the position of the Fermi level in the intrinsic semiconductor, and n_i is the intrinsic carrier density in the semiconductor.

The electron densities in the conduction band of two n-type semiconductors brought in contact with each other are related to the Fermi level and the energy level in the middle of the band gap by

$$n_1 = n_i e^{\frac{E_{F1} - E_{i1}}{kT}} \tag{1.1a}$$

$$n_2 = n_i e^{\frac{E_{F2} - E_{i2}}{kT}} \tag{1.1b}$$

Under equilibrium conditions the Fermi levels of two semiconductors brought in contact coincide, therefore

$$E_{F1} = E_{F2} \tag{1.2}$$

From (1.1a) and (1.1b) it follows that

$$E_{F1} = E_{i1} + kT \ln \frac{n_1}{n_i}$$

$$E_{F2} = E_{i2} + kT \ln \frac{n_2}{n_i}$$

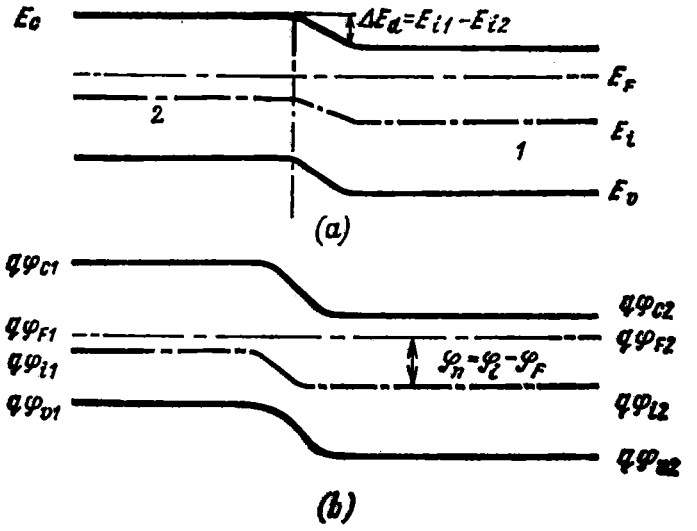


Fig. 1.1. Contact potential difference developed across a junction between two semiconductors with electron densities n_1 and n_2, respectively
(a) energy band diagram; (b) potential distribution diagram

Using equality (1.2) we may get

$$\Delta E_d = E_{i1} - E_{i2} = kT \ln \frac{n_2}{n_1} \tag{1.3}$$

Equation (1.3) shows that a change in the carrier density from n_1 to n_2 causes the energy bands to bend (tilt) by a value of ΔE_d.

The energy-band diagram of a contact between two semiconductors is shown in Fig. 1.1. Since the operation of semiconductor devices is commonly assessed through the voltage-current relationship, the energy diagrams are often replaced by the potential distribution diagrams. In this case the Fermi level is expressed in terms of Fermi potential φ_F:

$$E_F = q\varphi_F$$

Likewise, the energy levels at the lower edge (bottom) of the conduction band, at the upper edge (top) of the valence band, and in the middle of the band gap are given in terms of their potentials:

$$E_c = -q\varphi_c, \quad E_v = -q\varphi_v$$

$$E_i = \frac{E_c - E_v}{2} = q\frac{\varphi_c - \varphi_v}{2} = q\varphi_i$$

The quantity $\varphi_n = (E_i - E_F)/2q$ is called the macroscopic electrostatic potential of a semiconductor.

The electron and hole densities are related to these potentials by

$$n = n_i e^{-q\frac{\varphi_i - \varphi_F}{kT}}$$

$$p = n_i e^{-q\frac{\varphi_F - \varphi_i}{kT}}$$

Electron densities n_1 and n_2 (1.1a, 1.1b) are expressed in terms of the Fermi potentials in the following manner:

$$n_1 = n_i e^{-q\frac{\varphi_{i1} - \varphi_{F1}}{kT}}, \quad n_2 = n_i e^{-q\frac{\varphi_{i2} - \varphi_{F2}}{kT}} \tag{1.4a}$$

A potential difference $\Delta\varphi_d$ thus develops across the region where the electron density changes from n_1 to n_2.

Indeed, from (1.4a) it follows that

$$\varphi_{F1} = \varphi_{i1} + \frac{kT}{q}\ln\frac{n_1}{n_i}, \quad \varphi_{F2} = \varphi_{i2} + \frac{kT}{q}\ln\frac{n_2}{n_i} \tag{1.4b}$$

Since the Fermi potentials of the two semiconductors in contact coincide, we obtain from (1.4b)

$$\Delta\varphi_d = \varphi_{i1} - \varphi_{i2} = \frac{kT}{q}\ln\frac{n_2}{n_1} \tag{1.5}$$

This potential difference corresponds to the bend of the energy band edges on the potential distribution diagram of Fig. 1.1*b*. In this case the potentials at the bottom of the conduction band, at the top of the valence band, and in the middle of the band gap, φ_c, φ_v, and φ_i, respectively, are equal since the bends of the corresponding band edges are identical

$$\Delta\varphi_d = \Delta\varphi_c = \Delta\varphi_v = \frac{kT}{q}\ln\frac{n_2}{n_1}$$

The energy barrier ΔE_d and potential barrier $\Delta\varphi_d$ are formed across the region (boundary layer) with the changing electron density due to the diffusion of electrons from the semiconductor with a higher electron density to the semiconductor with a lower electron density. In the vicinity of the boundary layer there develops a positive space charge of ionized donor atoms (in the semiconductor with a higher electron density) and a negative space charge of electrons (in the semiconductor with a lower electron density).

The electric double layer so formed is being depleted of electrons, giving rise to an electric field and a potential difference $\Delta\varphi_d$, defined as the contact potential difference. If we apply a direct voltage to the contact so that semiconductor *2* (Fig. 1.1*a*) with a higher value of φ_F charges positively, the contact will have a high resistance.

The electrons flowing under the action of the applied voltage to semiconductor *2* must overcome an energy barrier equal to $q\Delta\varphi_d$. Should we reverse the current flow, the barrier to the electrons will be removed and the resistance of the contact will become small.

Thus, owing to the barrier the magnitude of the current flowing through the contact will vary significantly when the sign of the applied voltage is changed. This phenomenon is called rectification. The rectification effect is generally insignificant at the contact between semiconductors of the same type. The contact between two *n*-type semiconductors of different conductivities is referred to as n-n^+-junction. The sign$^+$ denotes the region of a higher conductivity.

A similar barrier is also formed at the boundary between *p*-type semiconductors of different conductivities. This contact is called p-p^+-junction.

The contact potential difference across n-n^+- and p-p^+-junctions is usually small.

Now consider how the potential difference develops at a semiconductor-to-metal contact. Its magnitude depends on the work functions of the contacting materials. The work function is the energy required to remove an electron from the surface of a solid into vacuum. When two bodies having different work functions are brought in contact, the electrons escaping from the body of a lower work function into the body of a higher work function set up a contact potential difference numerically equal to the difference in the work functions.

The work function of a semiconductor and a metal is counted off from the Fermi level and therefore it is higher for an electron in the valence band than for an electron lying on the impurity level or in the conduction band. This is why the work function of a *p*-type semiconductor, silicon, for example, is higher than the work function of an *n*-type semiconductor. The work function may vary over a wide range depending on the doping level of a semiconductor (Fig. 1.2*a*).

Besides, the work function of a metal or a semiconductor is materially affected by its surface state. For example, depositing a thinnest film of potassium on a platinum plate changes the work function of platinum from 5.22 to 1.62 eV. It will be worth noticing that the work function of potassium proper is 2.26 eV.

Table 1.1

Values of Work Functions for Metals and Semiconductors

Material	K	Cs	Ag	Au	Zn	Cd	Al	Tl	Ge	Si	Se	CdS	K on Pt	Pt
$q\varphi$, eV	2.26	1.93	4.73	4.76	4.25	4.08	4.25	3.92	4.55	4.3	4.89	4.58	1.62	5.22

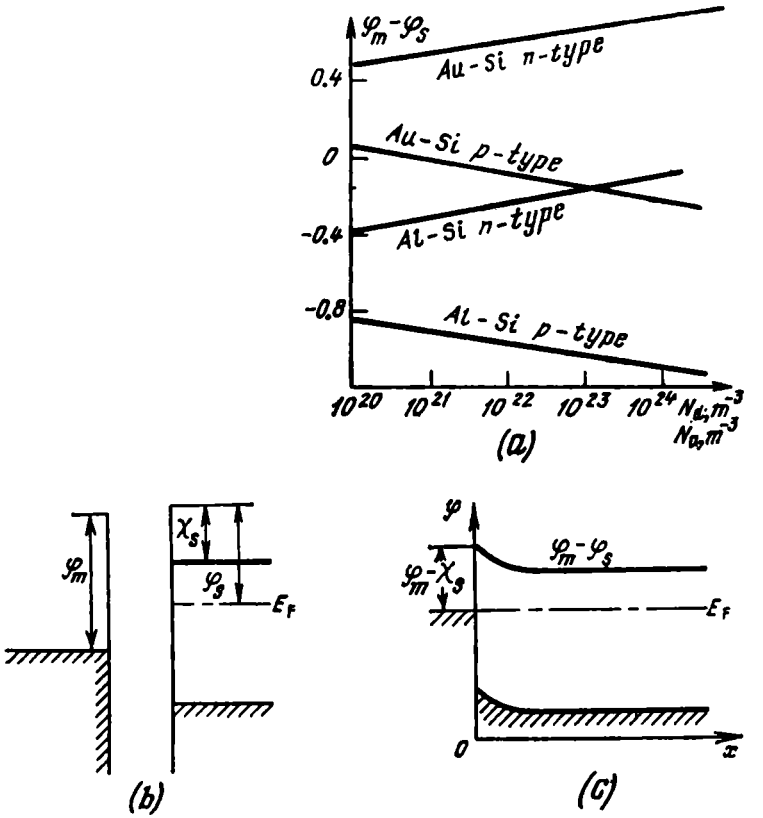


Fig. 1.2. Difference in work functions between metal and semiconductor *versus* impurity concentration in the semiconductor (*a*), and energy-band diagrams for metal and semiconductor (*b*) and for semiconductor-metal contact (*c*)

The work functions of some metals and semiconductors are given in Table 1.1 which shows that the potential difference across a semiconductor metal contact can be sufficiently high and the rectification effect associated therewith rather significant.

1.2. CURRENT RECTIFICATION AT SEMICONDUCTOR-METAL CONTACT

We may single out two groups of phenomena which account for the rectification at a metal-semiconductor contact:

(1) difference in the work functions between a metal and a semiconductor;

(2) presence of surface states creating intrinsic potential barriers.

This Section deals only with the potential barriers resulting from the difference between the work functions of a metal and a semiconductor. The surface effects will be discussed in the next Chapter.

Consider a contact between an n-type semiconductor and a metal whose work function is higher than that of the semiconductor, i.e., $\varphi_m > \varphi_S$.

Figure 1.2*b* shows the energy-band diagrams for a metal characterized by work function φ_m and a semiconductor having electron affinity χ_S and work function φ_S. The electron affinity is the energy of capturing an electron by a neutral atom. The electron affinity of some metals and oxygen is given in Table 1.2.

Table 1.2

Electron Affinity of Some Metals and Oxygen

Atom	O	Cu	Al	Au
χ_S, eV	3.0	1.0	1.1	2.4

The electrons of the semiconductor with a lower work function move into the metal, thus producing a positive space charge in the semiconductor until the electric field of the double layer at the contact starts impeding further electron transfer. This will result in the development of a potential barrier $\Delta\varphi_d = \varphi_m - \varphi_S$. Fig. 1.2*c* shows the energy-band diagram of the contact in equilibrium.

The width of the space charge layer in the metal is small due to the high density of electrons and may be disregarded. The space charge region in the semiconductor from which the electrons have escaped depends on the depth of penetration of the field into the semiconductor and is called the depletion layer whose width δ may be calculated with the Poisson equation:

$$\varepsilon\varepsilon_0 \frac{\partial^2\varphi(x)}{\partial x^2} = \rho(x) \tag{1.6}$$

where $\rho(x)$ is the space charge density, ε is the relative permittivity of the material, ε_0 is the dielectrical permittivity of vacuum, $\varphi(x)$ is the potential across section x, where x is the coordinate counted off from the contact and into the bulk of the semiconductor.

The positive charge of ionized impurity atoms in an n-type semiconductor is balanced by the negative charge of free electrons and the resultant charge is therefore zero:

$$\rho = q(N_d - n) = 0$$

If we assume that all the impurity atoms in the semiconductor are ionized at room temperature, i.e., $n = N_d$, and there are no free

electrons in the depletion layer, then the space charge density ρ in this layer equals $-qN_d$. The space charge density in the depletion layer of the semiconductor is considered to be invariable and equal to zero when $x = \delta$.

In solving equation (1.6) we assume that with the transition from the surface to the bulk of the semiconductor the change in the potential varies from $\Delta\varphi_d$ to 0, while at a distance $x \geqslant \delta$ no change in the potential is observed, which means that there is no electric field in the bulk of the semiconductor.

Based on the above assumptions, the boundary conditions for equation (1.6) may be written in the following form:

$$\varphi|_{x=0} = \Delta\varphi_d, \quad E|_{x=\delta} = -\frac{\partial\varphi}{\partial x}\bigg|_{x=\delta} = 0 \tag{1.7}$$

Integrating equation (1.6) gives

$$E = \frac{qN_d}{\varepsilon\varepsilon_0}(\delta - x), \quad \varphi = \frac{qN_d}{2\varepsilon\varepsilon_0}(\delta - x)^2 \tag{1.8}$$

Hence, the width of the depletion layer

$$\delta = \left(\frac{2\varepsilon\varepsilon_0\Delta\varphi_d}{qN_d}\right)^{1/2} \tag{1.9}$$

It is seen from (1.9) that the width δ of the depletion layer depends on the concentration of impurities and the permittivity of the semiconductor. With $N_d = 10^{14}$ cm^{-3}, $\varepsilon = 16$, $\Delta\varphi_d = 1$V, and $\varepsilon_0 = 8.85 \times 10^{-10}$ F/cm, $\delta \approx 4 \times 10^{-4}$ cm.

With an increase of the impurity concentration in the semiconductor, the width δ of the depletion layer sharply reduces: at $N_d = 10^{15}$ cm^{-3} it equals 10^{-5} cm and at $N_d = 10^{20}$ cm^{-3} it drops to 10^{-7} cm.

Thus, at the contact between the n-type semiconductor and the metal with a work function $\varphi_m > \varphi_S$ a barrier layer sets in whose width depends on the value of $\Delta\varphi_d$.

If $\varphi_m < \varphi_S$, the electrons will pass from the metal to the semiconductor forming a layer enriched in electrons. No potential barrier appears at the contact in this case and the electrons flowing from the boundary region easily penetrate into the bulk of the semiconductor. Such contacts are called ohmic since the currents passing through them obey Ohm's law. When a p-type semiconductor is brought in contact with a metal whose work function $\varphi_m > \varphi_S$, an ohmic contact is formed, while at $\varphi_m < \varphi_S$ a blocking contact is formed.

Four possible variants of contact between a semiconductor and a metal are shown in Fig. 1.3.

We shall now discuss the rectification effect that occurs at a semiconductor-metal contact. Let us apply an external voltage V to the

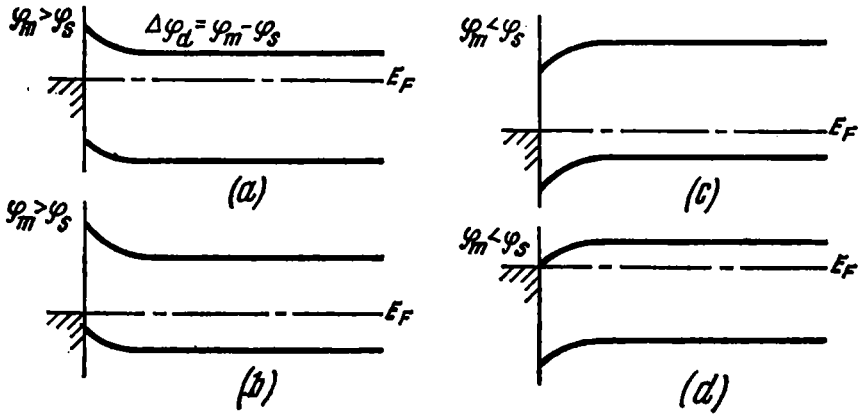


Fig. 1.3. Formation of depletion (blocking) layers (*a*, *c*) and enriched layers (*b*, *d*) at *n*-type semiconductor-metal junction and a *p*-type semiconductor-metal junction as a function of the difference between work functions (*b* and *c* refer to *p*-type semiconductors)

contact with the metal connected to the negative and the *n*-type semiconductor to the positive terminal of the battery. Since the resistance of the space charge layer is much higher than that of the semiconductor throughout its thickness, all the applied voltage drops across the contact. The height of the potential barrier at the junction between the semiconductor and the metal rises, therefore, to an amount $\Delta\varphi_d^* = \Delta\varphi_d + V$ as the external field coincides in sign with the field at the contact and strengthens the latter.

As a result, the electrons in the semiconductor flow from the boundary into the bulk of the semiconductor so that the width of the space charge layer increases above the equilibrium value. The voltage of this sign is called barrier or reverse voltage (Fig. 1.4*b*). If we apply a voltage of opposite sign to the contact, i.e., connect the metal with the positive terminal (Fig. 1.4*c*), the external field will reduce the height of the potential barrier $\Delta\varphi_d = \Delta\varphi_d - V$, in other

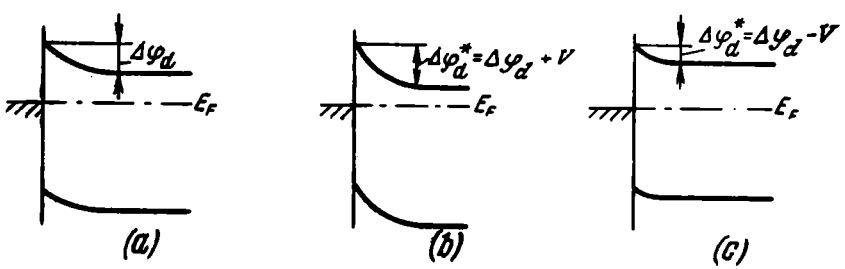


Fig. 1.4. Energy band model of a semiconductor-metal contact
(*a*) in equilibrium; (*b*) reverse voltage applied; (*c*) forward voltage applied

words, add to the transfer of electrons from the semiconductor to the metal and diminish the width of the barrier, decreasing thereby the barrier resistance and providing for the flow of a large current through the contact. This direction of the current flow is termed forward.

Thus, if a space charge layer is formed at the semiconductor-metal contact, the latter displays the properties of an electric current rectifier because the contact resistance proves high with one sign of the voltage applied and low, with the opposite sign.

With the external voltage applied across the contact, the width of the space charge layer is determined from the following relation:

$$\delta = \left[\frac{2\varepsilon\varepsilon_0 (\Delta\varphi_d \pm V)}{qN_d}\right]^{1/2} \tag{1.10}$$

1.3. RECTIFICATION AT METAL-SEMICONDUCTOR CONTACT

The current flow through the potential barrier of a contact depends on the width of the space charge region.

When the width of this region exceeds the free path of electrons, the charge carriers passing over the space charge layer at the forward voltage applied suffer numerous collisions and overcome them by diffusing across the layer.

If the width of the space charge layer is smaller than the free path of electrons, the electrons present in the depletion layer are not subjected to collisions. The electrons whose kinetic energy is sufficient to overcome the potential barrier of the contact can move from the semiconductor to the metal.

In conformity with these mechanisms of electron transfer across the contact the current-voltage relationship is described by either the diffusion or the diode theory.

Schottky Diffusion Theory

According to the diffusion theory, the current-voltage relationship or current-voltage characteristic of a contact is determined on the assumption that the electron flow consists of two components, namely, a diffusion and a conduction current. The diffusion current density J_{dif} depends on the electron density gradient in the space charge layer and the diffusion factor. It may thus be described by the expression

$$J_{dif} = qD \operatorname{grad} n$$

The conduction current density J_0 is described by Ohm's law and is written as

$$J_0 = qn\mu E$$

where μ is the charge carrier mobility.

The resultant net current through the contact comprises the sum of the above currents[1]

$$J = qn\mu E + qD \text{ grad } n \tag{1.11}$$

It is assumed that the flow of minority carriers may be disregarded.

The electron diffusion factor D is related to the charge carrier mobility μ and thermal potential kT/q by the Einstein relation:

$$D_{p,n} = \mu \frac{kT}{q}$$

Suppose now that the electric field and electron density distribution vary only in the direction of x, then (1.11) takes on the form

$$J(x) = qn\mu E(x) + \mu kT \frac{dn}{dx} \tag{1.12}$$

$E(x)$ is defined here by the field of the contact potential difference and the external voltage applied, therefore

$$E(x) = -\frac{d\varphi}{dx} = \frac{d(\varphi_d + V)}{dx}$$

Equation (1.12) can be rewritten in the form

$$\frac{dn}{dx} - \frac{n}{kT}\frac{d\varphi}{dx} = \frac{J}{qD} \tag{1.13}$$

Since all the voltage applied to the contact may be assumed to drop off across the space charge region, then at point $x = 0$ potential $\varphi_d^* = \Delta\varphi_d + V$, and at point $x = \delta$ the potential goes to zero. The carrier density at point $x = 0$ is therefore

$$n|_{x=0} = n_b e^{\frac{-q(\Delta\varphi_d + V)}{kT}}$$

where n_b is the equilibrium density of carriers beyond the space charge region.

Based on these assumptions, the boundary conditions for equation (1.13) may be written as

$$\varphi|_{x=0} = \Delta\varphi_d + V, \quad n|_{x=0} = n_b e^{\frac{-q(\Delta\varphi_d + V)}{kT}} \tag{1.14}$$

$$\varphi|_{x=\delta} = 0, \quad n|_{x=\delta} = n_b \tag{1.15}$$

The solution of the equation is

$$J = \frac{qDn_b}{kT}\left(\frac{d\varphi}{dx}\right)_{x=0} e^{\frac{-q\Delta\varphi_d}{kT}} \left(e^{-q\frac{V}{kT}} - 1\right) \tag{1.16}$$

It follows from (1.14) that with no external circuit

$$n|_{x=0} = n_b e^{-q\frac{\Delta\varphi}{kT}}$$

[1] In Chapters 1 and 2 the term current will stand for current density J.

and then, designating the equilibrium density of electrons on the surface as $n_s = n|_{x=0} = n_b e^{\frac{-q\Delta\varphi_d}{kT}}$ and assuming that at $x = 0$ the $d\varphi/dx$ may be approximated by the equation

$$-\frac{d\varphi}{dx}\bigg|_{x=0} = E_0 = \frac{\Delta\varphi_d + V}{\delta}$$

we obtain from (1.16)

$$J = qn_s\mu E_0 (e^{\frac{-qV}{kT}} - 1) \tag{1.17}$$

Expression (1.17) describes the current-voltage characteristic of a semiconductor-metal contact.

If the external voltage is so applied to the contact that the semiconductor is connected to the negative terminal of the battery ($V < 0$), the current through the contact grows rapidly by the law

$$J = J_c e^{\frac{qV}{kT}}$$

This is a forward direction of current. In the reverse direction ($V > 0$), current $J \approx -qn_s\mu E_0$, i.e., rises linearly with an increase in the applied voltage.

Diode Theory of Rectification

We may neglect the diffusion component when calculating the current through thin space charge layers, and may consider therefore that only those electrons move from the semiconductor to the metal the kinetic energy of which is sufficient to pass over the contact potential barrier of height $q(\Delta\varphi_d + V)$. For these electrons the following relation must be observed

$$\frac{m_0 v^2}{2} \geqslant q(\Delta\varphi_d + V) \tag{1.18}$$

where $m_0 v^2/2$ is the kinetic energy of an electron.

The number of these electrons passing through a semiconductor area of one square centimetre during one second is defined by the equation

$$n_1 = \frac{1}{2} n_b \bar{v} e^{-\frac{q(\Delta\varphi_d + V)}{kT}} \tag{1.19}$$

where $\bar{v}$ is the mean electron velocity equal to $(m^*/2\pi kT)^{1/2}$, and m^* is the effective mass of electron.

The current-voltage characteristic of the contact may be obtained if we regard the current through the contact as a sum of two cur-

rents, namely, the current flowing from the semiconductor into the metal and that from the metal into the semiconductor. Let the current flowing from the semiconductor into the metal be the product of a number of electrons capable of leaving the semiconductor by an electron charge

$$J_1 = qn_1 = \frac{1}{2} qn_b \bar{v} e^{-\frac{q\Delta\varphi_d}{kT}} e^{-\frac{qV}{kT}} \tag{1.20}$$

Since the electron density in the metal is very high ($n = 10^{23}$ cm^{-3}), the external voltage is not able to change the position of the Fermi level in the metal, and therefore the current flowing from the metal into the semiconductor is independent of the external voltage and is determined only by the work function of the metal.

This current may be written in the following form:

$$J_2 = \frac{1}{2} q\bar{v} n_s \tag{1.21}$$

The resultant current is

$$J = J_1 - J_2 = \frac{1}{2} qn_b e^{-\frac{q\Delta\varphi_d}{kT}} (e^{-\frac{qV}{kT}} - 1)$$

or considering that

$$J_0 = \frac{1}{2} qn_b \bar{v} e^{-\frac{q\Delta\varphi_d}{kT}} \tag{1.22}$$

we obtain

$$J = J_0 (e^{-\frac{qV}{kT}} - 1) \tag{1.23}$$

J_0 is termed the thermal current as its value is defined by the semiconductor parameters and temperature. Equation (1.23) describes the current-voltage characteristic of the contact.

Connecting the semiconductor to the negative side of the voltage source causes the current through the contact to rise exponentially (forward direction) due to the fact that the height of the potential barrier, $\Delta\varphi_d^* = \Delta\varphi_d - V$, diminishes and an increasing number of electrons are being swept from the semiconductor into the metal.

For the reverse direction ($V > 0$), when $V \geqslant 3\ kT/q$, i.e., $V \geqslant 78$ mV, the term $e^{\frac{qV}{kT}} \ll 1$ and the current is determined only by the work function and is independent of the magnitude of the voltage.

When comparing equations (1.17) and (1.23), notice that current J_0 in (1.23) does not depend on the voltage, and for this reason the reverse current curve (Fig. 1.5) calculated by the diode theory has a saturation region.

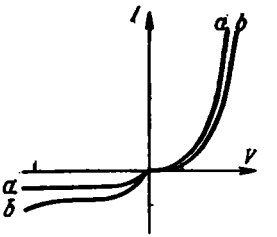


Fig. 1.5. Current-voltage characteristics for a semiconductor-metal contact plotted according to diode theory (*a*) and diffusion theory (*b*)

It should be noted that the concepts of thin and thick space charge layers for which the above given theories hold true are not strict. Indeed, as the electron free path λ is associated with the mobility of carriers, the depletion layer width δ in semiconductors with a low carrier mobility proves thick because $\delta > \lambda$ and thin in semiconductors with a high carrier mobility. Thus, the current-voltage characteristics of diodes made of selenium and cuprous oxide substances with a low mobility can be adequately described through the use of the Schottky theory, while the rectification effect at the contact between a metal and germanium or silicon is more amenable to the description by the diode theory.

There is however a discrepancy between theoretical calculations and experimental results. Thus, the current-voltage characteristics determined from formulas (1.17) and (1.23) display comparatively low rectification ratios (ratios of the forward current to the reverse current at a constant voltage), though actual rectifiers have considerably higher ratios reaching several hundreds and thousands.

High rectification ratios of some semiconductor-metal contacts are associated with the formation under the metal of a thin *p*-type region in an *n*-type semiconductor, i.e., the formation of a *p-n* junction in the semiconductor.

The reverse current calculated on the basis of these theories proves to be considerably lower than that actually observed. High values of the reverse current through a semiconductor-metal contact are attributed to minority carriers.

Indeed, formulas (1.17) and (1.23) have been derived on the assumption that a semiconductor has only majority carriers (electrons in an *n*-type semiconductor and holes in a *p*-type semiconductor). In actual fact a certain equilibrium quantity of minority carriers is always present in a semiconductor (holes in an *n*-type and electrons in a *p*-type semiconductor).

Since the electric field of a contact energized by the reverse voltage repels free majority carriers from the contact and pulls over minority carriers to the depletion layer, the current flowing through the potential barrier turns out to be larger than the value calculated with formulas (1.17) and (1.23).

Review Questions

1. Find the bend of the energy bands near the surface of a semiconductor, if its surface conductivity is 10 times its bulk conductivity.
2. What hampers the inversion of conduction in low-resistivity semiconductors?
3. Why cannot the contact between two metals with different work functions be used as a rectifier?
4. Calculate the equilibrium width of the depletion layer in a semiconductor with $N_d = 10^{+17}$ cm^{-3}, $\varphi_d = 0.25$ V, and $\varepsilon = 16$.
5. Why are the results of calculation by the diffusion theory closer to the experimental results for a selenium diode than those obtained from the use of the diode theory?
6. Why does the current-voltage characteristic plotted by the use of the diffusion theory not display a reverse current saturation portion?
7. Why do the Fermi levels of the bodies in contact coincide?

Chapter Two

ELECTRON-HOLE JUNCTION

2.1. TYPES OF *p-n* JUNCTIONS. PHYSICAL MODEL OF A *p-n* JUNCTION

A junction between two regions of a semiconductor one of which has electron and the other hole conduction is termed *p-n* junction.

For a *p-n* junction to be formed, an exchange of carriers between an *n*-type and a *p*-type semiconductor region should take place, which means that the distance between the atoms in the *n*- and *p*-regions of the semiconductor must not exceed the value equal to the lattice constant, in other words, it must be only several angstroms (1 Å = 10^{-8} cm). This condition is met when a semiconductor crystal of one conductivity type (*n*-type, for example) is doped to form a region of another conductivity type (*p*-type, for example). Consider a contact between an *n*-type and a *p*-type semiconductor (Fig. 2.1). Suppose donor impurities predominate on the left of the boundary (*n*-type semiconductor) and acceptor impurities on the right of the boundary (*p*-type semiconductor).

The ionization energies of the impurities used for doping germanium and silicon reach 0.01-0.05 eV and at room temperature the impurities may be considered fully ionized. That is why the densities of mobile charge carriers coincide with those of impurity atoms, i.e., $n \simeq N_d$, and $p \simeq N_a$. Since on both sides of the boundary plane ($x = 0$) there are two regions, one of which has a prevailing concentration of free electrons and the other of holes, the electrons will diffuse across the boundary from the *n*-type into *p*-type region, and the holes, from the *p*-type into *n*-type region.

If the electrons and holes were electrically neutral, this process would continue until equilibrium of the electron and hole densities in the semiconductor is reached. In reality the diffusion of charge carriers ceases by far earlier because the electrons and holes are charged.

While diffusing across the junction, the electrons and holes leave donor- and acceptor-impurity ions in the crystal lattice sites, which form fixed positive and negative charges respectively in the *n*-type and *p*-type regions. This upsets the electrical neutrality of the semiconductor near the boundary, which leads to the formation in the boundary region of an electrical double layer consisting of holes and positively charged ions in the *n*-type region, electrons and negatively charged ions in the *p*-type region. This double layer is called the space-charge region or *p-n* junction region.

The direction of the double layer field is such that it impedes further diffusion of electrons to the *p*-region and holes to the *n*-region.

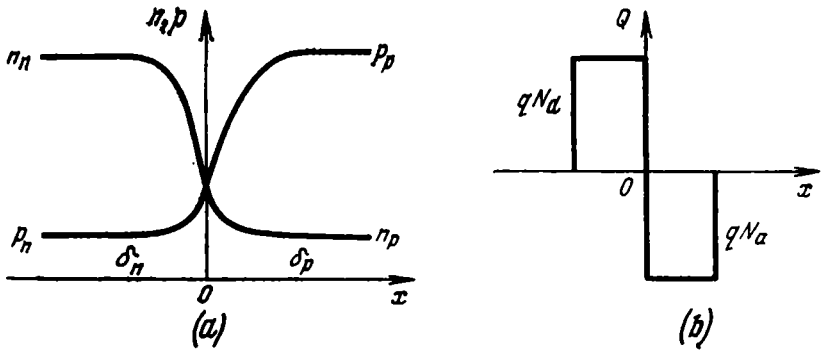


Fig. 2.1. Equilibrium distribution of charge carrier density (*a*) and space charge density (*b*) on both sides of the boundary between *p*- and *n*-regions

The same field draws into the junction the electrons present in the *p*-region and holes present in the *n*-region.

Suppose the electron and hole densities in the *n*- and *p*-regions, respectively, are the same and equal to 10^{17} cm^{-3}. The *n*- and *p*-regions also have minority carriers, namely, holes in the *n*-region and electrons in the *p*-region, their densities being related to those of majority carriers by

$$np = n_i^2 \tag{2.1}$$

The densities of these minority carriers are

$$n_p = \frac{n_i^2}{p_p}, \qquad p_n = \frac{n_i^2}{n_p}$$

Assuming that $n_i = 10^{13}$ cm^{-3}, we find

$$n_p = p_n = \frac{10^{26}}{10^{17}} = 10^9 \text{ cm}^{-3}$$

Now let it be shown that a potential barrier $\Delta\varphi_d$ sets in at the boundary between the *p*- and *n*-regions. The electron density in the *n*-region and the hole density in the *p*-region are expressed in terms of the Fermi potentials φ_F and potentials φ_i in the middle of the band gap width:

$$n_n = n_i e^{-q\frac{\varphi_{in}-\varphi_F}{kT}}; \quad p_p = n_i e^{-q\frac{\varphi_F-\varphi_{ip}}{kT}}$$

or

$$\varphi_{in} + \varphi_F = \frac{kT}{q}\ln\frac{n_n}{n_i}, \quad \varphi_{ip} - \varphi_F = \frac{kT}{q}\ln\frac{p_n}{n_i} \tag{2.2}$$

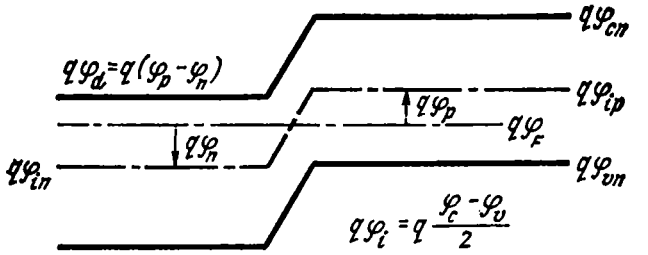


Fig. 2.2. Energy-band diagram of a *p-n* junction

Since the Fermi potentials of the semiconductors brought in contact coincide, it follows from (2.2) that

$$\Delta\varphi_d = \varphi_{ip} - \varphi_{in} = \frac{kT}{q} \ln \frac{p_p n_p}{n_i^2} \tag{2.3a}$$

Considering (2.1), expression (2.3a) takes the form

$$\Delta\varphi_d = \frac{kT}{q} \ln \frac{n_n}{n_p} = \frac{kT}{q} \ln \frac{p_p}{p_n} \tag{2.3b}$$

For germanium with $n_i = 10^{13}$ cm^{-3}, $n_n = 10^{19}$ cm^{-3} and $n_p = 10^7$ cm^{-3},

$$\Delta\varphi_d = 28\frac{kT}{q} \simeq 0.7 \text{ V}$$

Note that the value of the contact potential barrier may be expressed in terms of electrostatic potentials φ_p and φ_n of the *p*- and *n*-regions, respectively

$$\varphi_p = \varphi_{ip} - \varphi_F; \quad \varphi_n = \varphi_{in} - \varphi_F \tag{2.4a}$$

We find from (2.3)

$$\Delta\varphi_d = \varphi_{ip} - \varphi_{in} = \varphi_p - \varphi_n \tag{2.4b}$$

The energy-band diagram of a *p-n* junction is shown in Fig. 2.2.

Consider now the charge distribution in a *p-n* junction. Referring to Fig. 2.1*a*, the maximum depletion or the lowest density of mobile carriers ($n = p = n_i$) is observed at the boundary between the *p*-region and *n*-region, namely, across section $x = 0$, where the total charge of ionized donor and acceptor atoms and also of electrons and holes is equal to zero, so that

$$\rho = q(N_d + p - N_a - n) = 0$$

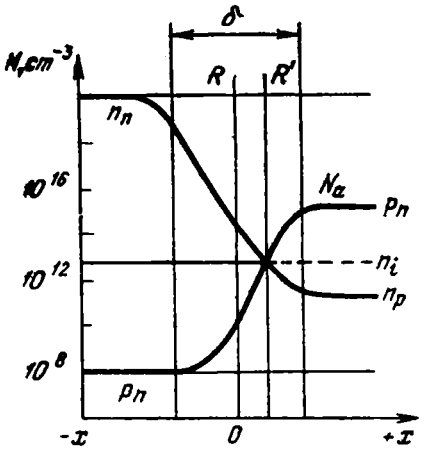


Fig. 2.3. Distribution of free charge carriers in an asymmetrical *p-n* junction

δ is the width of the space-charge region in the *p-n* junction

On both sides of section $x = 0$ the electrical neutrality is upset and the *n*-region has an uncompensated positive charge with the density

$$\rho = q\,(N_d - n)$$

and the *p*-region has an uncompensated negative charge with the density

$$\rho = -q\,(N_a - p)$$

The density of free carriers just near the boundary can be neglected when $x \simeq 0$. If that is the case, the charge densities in the *n*-region and *p*-region are

$$\rho_n = qN_d, \quad \rho_p = -qN_a$$

The density of free electrons in the *n*-region and that of holes in the *p*-region increase with the distance from the boundary. When $x = \pm\delta$ charges ρ_n and ρ_p drop to zero because in this case $N_d = n$ and $N_a = p$.

The quantity $\delta_n + \delta_p$ determines the width of the *p-n* junction. For simplicity, we assume that the charge does not vary and equals qN_d across the whole width of the *p-n* junction, i.e. when $x < \pm\delta$, and falls to zero when $x \geqslant \pm\delta$. The charge distribution is illustrated in Fig. 2.1*b*. The free carrier density distribution in this *p-n* junction is symmetrical about the n_i-level and the interface between the *p*- and *n*-regions.

The above considered junction whose electron density in the *n*-region is equal to that of holes in the *p*-region is called symmetrical. When the concentrations of impurities in the *p*- and *n*-regions

differ, the p-n junction is termed asymmetrical. For such a junction, the curves of the carrier density distribution in the n- and p-regions are symmetrical about the n_i-level and asymmetrical about the boundary between the p- and n-regions.

In the asymmetrical p-n junction the plane where $n_p = n_n = n_i$ is shifted from the boundary (R in Fig. 2.3) towards a lower density of charge carriers (plane R'). This shift is attributed to the fact that the thickness (width) of the space-charge region is greater in that semiconductor portion where the carrier density is smaller. For this reason, when the concentrations of impurities in the p-region and n-region differ considerably, for example, when $N_a \ll N_d$, practically all the space-charge layer lies in one of the regions (the p-region in this case).

One can distinguish graded or abrupt junctions depending on the concentration gradient of impurities near the p-n boundary. The junction in which the width of the region with a changing impurity density is much less than the width of the space-charge layer is identified as an abrupt junction, and the junction in which the width of the region with a changing impurity density is comparable with that of the space-charge layer is referred to as a graded junction.

For the space-charge layer to be widened, a semiconductor layer with an intrinsic (least possible) conduction (i-layer) is sometimes formed between the n- and p-regions. The junction of this type is known as a p-i-n junction.

2.2. CALCULATION OF THE WIDTH OF SPACE-CHARGE LAYER IN THE p-n JUNCTION

Abrupt p-n Junction

To determine the width of a space-charge layer in an abrupt p-n junction recourse can be made to the fact that this layer has an uncompensated electric charge, electric field and a potential barrier at the p-n boundary. Beyond the region of the space charge the electric field, potential barrier and uncompensated charge are equal to zero. It may be assumed that the charge in the p-n junction is distributed in the manner as is shown in Fig. 2.1b, i.e., in the p-region $\rho = -qN_a$ when $0 \leqslant x \leqslant -\delta$, and in the n-region $\rho = qN_d$ when $0 \leqslant x < \delta$. Outside the p-n junction, i. e., when $x \geqslant \delta$, $E = 0$.

The distribution of electric field E and potential φ throughout the p-n junction is found from the Poisson equation

$$\frac{d^2\varphi}{dx^2} = \frac{\rho}{\varepsilon\varepsilon_0} \tag{2.5}$$

The boundary conditions for the equation are determined on the assumption that outside the p-n junction the electric field E

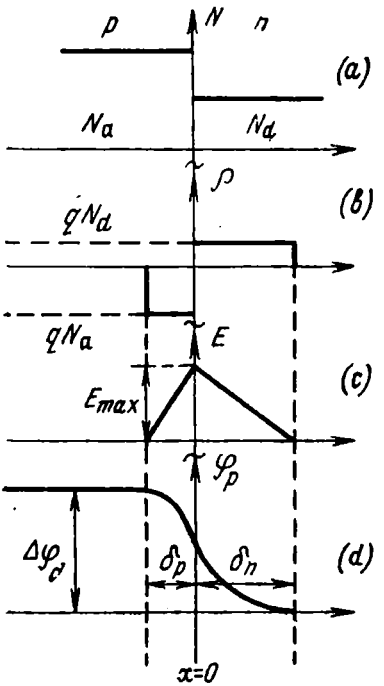


Fig. 2.4. Distribution of impurities (*a*), space charge (*b*), electric field strength (*c*), and potential (*d*) in an abrupt asymmetrical *p-n* junction

is zero and potential φ is equal to the electrostatic potential, i. e.,

$$E = 0 \quad \text{when} \quad x \geqslant \pm\delta \tag{2.6a}$$

$$\varphi = \varphi_n \quad \text{when} \quad x \geqslant \delta \tag{2.6b}$$

$$\varphi = \varphi_p \quad \text{when} \quad x \leqslant -\delta \tag{2.6c}$$

With boundary conditions (2.6a) integration of the Poisson equation yields:

for a *p*-region ($x < 0$)

$$E_p = -\frac{qN_a}{\varepsilon\varepsilon_0}(x + \delta_p) \tag{2.7a}$$

for an *n*-region ($x > 0$)

$$E_n = \frac{qN_d}{\varepsilon\varepsilon_0}(x - \delta_n) \tag{2.7b}$$

Integrating (2.7) with boundary conditions (2.6b) and (2.6c) we obtain the law of variation of potential $\varphi(x)$ across the *p*- and *n*-regions:

$$\varphi(x) - \varphi_p = -\frac{qN_a}{2\varepsilon\varepsilon_0}(x + \delta)^2 \tag{2.8}$$

$$\varphi(x) - \varphi_n = -\frac{qN_d}{2\varepsilon\varepsilon_0}(x - \delta)^2 \tag{2.9}$$

Expressions (2.8) and (2.9) show that the potential of the *p-n* junction varies parabolically. Variations in the magnitude of the electrostatic field and the potential are illustrated in Fig. 2.4.

With $x = 0$, the electric fields in the *n*- and *p*-regions must be equal and, therefore, by equating (2.7a) and (2.7b) we can obtain the relation of the space-charge layer width in the *n*-region to that in the *p*-region:

$$\frac{\delta_n}{\delta_p} = \frac{N_a}{N_d} \tag{2.10}$$

It follows from (2.10) that for a symmetrical junction the space-charge layer width in a *p*-region is equal to that in an *n*-region. For an asymmetrical junction, when $N_a \gg N_d$

$$\delta_p \ll \delta_n$$

and thus the total width of the space-charge layer

$$\delta_0 = \delta_n + \delta_p \simeq \delta_n \tag{2.11}$$

This means that all the space-charge region is located in a higher resistivity portion of a semiconductor.

Based on the condition of equality of potentials at the boundary between the n- and p-regions at point $x = 0$, it follows from (2.8 and 2.9), in view of (2.4a), that

$$\delta_0 = \sqrt{\frac{2\varepsilon\varepsilon_0\Delta\varphi_d}{q}\left(\frac{1}{N_a}+\frac{1}{N_d}\right)}$$

For an asymmetrical junction with $N_a \gg N_d$, this expression has the form

$$\delta_0 = \sqrt{2\varepsilon\varepsilon_0\Delta\varphi_d/qN_d} \qquad (2.12)$$

It is easy to see that the derived expression agrees with formula (1.5) which describes the potential barrier width at a semiconductor-to-metal contact.

Graded *p-n* Junction

When analysing a graded *p-n* junction, account is taken of the fact that the space charge in the *p-n* junction region does not depend on coordinate x.

The solution of the Poisson equation under the assumption of a linear dependence of the space-charge density on the coordinate, which is frequently supported in practice, leads to the following relation between the width of the space-charge layer and the impurity concentration:

$$\delta_0 = \left(\frac{12\Delta\varphi_d\varepsilon\varepsilon_0}{q}\right)^{1/3}\left(\frac{1}{N'_d}+\frac{1}{\sqrt{1/N'_aN'_d}}+\frac{1}{N'_d}\right)^{1/3} \qquad (2.13)$$

where N'_d is the donor density gradient in the *p-n* junction region, and N'_a is the acceptor density gradient in the *p-n* junction region.

The quantity $\Delta\varphi_d$ in (2.13) is the same for the graded junction and the abrupt junction and is calculated from formula (2.3).

In an asymmetrical graded junction with $N'_a \gg N'_d$, the width of the space-charge layer is given by

$$\delta_0 = \sqrt[3]{12\varepsilon\varepsilon_0\Delta\varphi_d/qN'_d} \qquad (2.14)$$

The notion of the abrupt and graded *p-n* junction is relative. In low-resistivity semiconductors the width of the space-charge layer found from formula (2.12) is less than 1 μm.

If the layer of varying impurity density is about 1 μm thick, computations must be done with the use of formulas for a graded *p-n* junction.

Higher-resistivity semiconductors with $\delta_0 > 5$μm but with the same thickness of the layer of varying impurity density can be calculated with high accuracy by using formulas for an abrupt *p-n* junction.

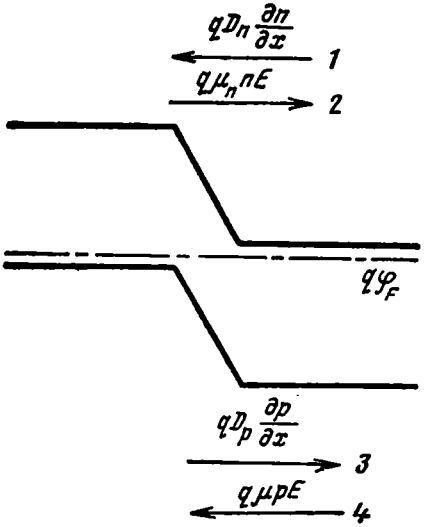


Fig. 2.5. Direction of flow of current components across a *p-n* junction
1—diffusion current of electrons; *2*—drift current of electrons; *3*—diffusion current of holes; *4*—drift current of holes

2.3. RECTIFICATION AT THE *p-n* JUNCTION

An electron density gradient present in the region of a *p-n* junction gives rise to an electron diffusion current whose density $J_{n\,dif} = -qD_n \frac{\partial n}{\partial x}$. This current is cancelled by an equally large electron conduction current $J_{n0} = qn\mu E$ due to the transfer of electrons from the *p*-region to the junction under the action of the electric field.

In a similar manner, from the *p*-region into the *n*-region there flows a hole diffusion current $J_{p\,dif} = qD_p \frac{\partial p}{\partial x}$ balanced by a hole conduction current (Fig. 2.5).

Now we estimate the values of these currents. Suppose $p_p = 10^{17}$ cm^{-3}, $n_n = 10^{15}$ cm^{-3}, $D_p = 44$ cm$^2 \cdot$s^{-1}, $D_n = 93$ cm$^2 \cdot$s^{-1} and $q = 1.6 \times 10^{-19}$ C.

The expression for the diffusion currents may be written in the following form:

$$J_{n\,dif} = qD_n \frac{\partial n}{\partial x} \simeq qD_n \frac{n_n - n_p}{\delta} \simeq qD_n \frac{n_n}{\delta}$$
$$J_{p\,dif} = qD_p \frac{\partial p}{\partial x} \simeq qD_p \frac{p_p - n_n}{\delta} \simeq qD_p \frac{p_p}{\delta} \qquad (2.15)$$

Considering that the space-charge layer width $\delta = 10^{-4}$ cm, we get

$$J_{n\,dif} = 75 \text{ A/cm}^2 = J_{n0}$$
$$J_{p\,dif} = 7 \times 10^3 \text{ A/cm}^2 = J_{p0}$$

In reality only a net current passes across the *p-n* junction, which is equal to the algebraic sum of current components $J_{n\,dif}$, J_{p0}, J_{n0} and $J_{p\,dif}$ and is many times less in magnitude than their arithmetic sum. This current does not exceed 1 A/cm^2 even when a forward bias is applied. Without any externally applied voltage, the net current of the four components is certainly equal to zero.

Now we apply an external voltage source to the *p-n* junction by connecting the positive terminal to the *p*-region and the negative terminal to the *n*-region. If we disregard a voltage drop across the

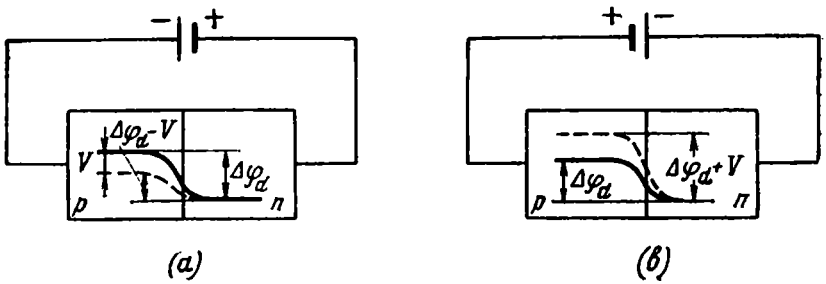


Fig. 2.6. Variations in the height of a potential barrier on applying a forward bias (*a*) and a reverse bias (*b*) on the *p-n* junction
Solid curve represents an equilibrium potential distribution

semiconductor bulk and assume that all the external voltage V is impressed across the *p-n* junction, the potential barrier $\Delta\varphi$ at the boundary between the *p*- and *n*-regions will change and equal

$$\Delta\varphi = \Delta\varphi_d - V$$

The decrease in the height of the barrier will result in an increase of electron- and hole-diffusion currents because their flow is now less impeded due to a weaker electric field at the contact. The hole and electron currents opposite in sign to the diffusion currents will diminish in this case whereby a certain forward current J will pass across the *p-n* junction.

A change in the height of the potential barrier at a *p-n* junction is illustrated in Fig. 2.6*a*. The solid and the dotted curves here are the plots of the potential distribution across the *p-n* junction in equilibrium and under applied bias V, respectively.

Substituting the new value of the barrier height $\Delta\varphi_d - V$ for $\Delta\varphi_d$ into (2.12), we obtain the width of the space-charge layer for an abrupt junction

$$\delta = \left[\frac{2\varepsilon\varepsilon_0(\Delta\varphi_d - V)}{qN_d}\right]^{1/2} = \delta_0\sqrt{\frac{\Delta\varphi_d - V}{\Delta\varphi_d}} \qquad (2.16)$$

where δ_0 is the width of the barrier in equilibrium, which is calculated from formula (2.12).

Thus, a forward voltage applied to the junction also decreases the width of the potential barrier, and with $\Delta\varphi_d = V$ the barrier width becomes equal to zero.

Since the equilibrium densities of carriers of the same type distributed on both sides of the barrier are interrelated by (2.3a), a change in the height of the barrier entails a change in the carrier density distribution. Let us write (2.3a) in the form $p_n = p_p e^{-q\Delta\varphi_d/kT}$. Replacing $\Delta\varphi_d$ by $\Delta\varphi = \Delta\varphi_d - V$ gives the hole density in the

n-region

$$p_n = p_p e^{\frac{-q(\Delta\varphi - V)}{kT}} = p_p e^{\frac{-q\Delta\varphi_d}{kT}} e^{\frac{qV}{kT}} \tag{2.17}$$

Bearing in mind that $p_p e^{-q\Delta\varphi_d/kT}$ is an equilibrium quantity of the hole density in the n-region in the absence of the external voltage, expression (2.17) takes the form

$$p_n = p_{n0} e^{\frac{qV}{kT}} \tag{2.18a}$$

where

$$p_{n0} = p_p e^{\frac{-q\Delta\varphi_d}{kT}}$$

Likewise, the electron density in the p-region

$$n_p = n_{p0} e^{qV/kT} \tag{2.18b}$$

It is obvious from (2.18) that as the forward voltage increases, the electron density in the p-region and hole density in the n-region grow exponentially. Such a transfer of minority carriers across the p-n junction is called injection. As the injected minority carriers move away from the p-n junction, their density diminishes due to recombination and at a distance x from the p-n junction it is equal to

$$p = p_{n0} + p_n(0)\, e^{-x/L_p} \tag{2.19}$$

where p_{n0} is the equilibrium density of holes in the n-region determined on condition that $n_n p_n = n_i^2$; $p_n(0)$ is the nonequilibrium hole density in the plane $x = 0$, found from (2.18a); and L_p is the hole diffusion length, i.e., a distance at which the hole density diminishes by a factor of e. The relation $p = f(x)$ is illustrated in Fig. 2.7.

Let us apply a reverse voltage to the p-n junction by connecting the positive terminal of the energy source to the n-region and the negative terminal, to the p-region of a semiconductor.

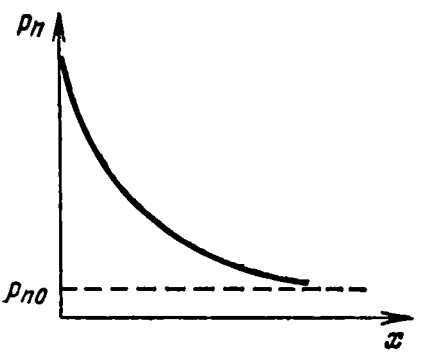


Fig. 2.7. Distribution of injected holes over the depth of an n-region in the p-n junction

The bias voltage of this polarity causes the barrier at the *p-n* boundary to increase to

$$\Delta\varphi = \Delta\varphi_d + V$$

This case of potential distribution across a *p-n* junction is represented on Fig. 2.6*b*.

An increasing potential barrier retards the flow of electron- and hole-diffusion currents and, besides, a stronger field at the contact reduces the number of majority carriers in the space-charge region, with the result that the depletion region at the *p-n* junction widens.

The width of the depletion region is given by

$$\delta = \delta_0 \sqrt{\frac{\Delta\varphi_d + V}{\Delta\varphi_d}}$$

When $V \gg \Delta\varphi_d$

$$\delta = \delta_0 \sqrt{\frac{V}{\Delta\varphi_d}} \tag{2.20}$$

2.4. CALCULATION OF THE CURRENT-VOLTAGE CHARACTERISTIC OF A *p-n* JUNCTION

In calculating the current-voltage characteristic we assume that there is no electric field beyond the *p-n* junction region and that the injected electrons and holes diffuse from the *p-n* junction into the bulk of the semiconductor. This means that we can take into account only diffusion components of the current.

Electron and hole diffusion is described by continuity equations which define the conditions of dynamic equilibrium of the minority-carrier density in a semiconductor. For holes in the *n*-region and electrons in the *p*-region they have the form

$$D_p \frac{\partial^2 p}{\partial x^2} = \frac{p - p_{n0}}{\tau} \tag{2.21a}$$

$$D_n \frac{\partial^2 n}{\partial x^2} = \frac{n - n_{p0}}{\tau} \tag{2.21b}$$

where D_n and D_p are the diffusion coefficients for holes and electrons, and τ is the electron and hole lifetime.

Equations (2.21) may be written in a somewhat different form

$$\frac{\partial^2 p}{\partial x^2} = \frac{p - p_{n0}}{L_p^2} \tag{2.22a}$$

$$\frac{\partial^2 n}{\partial x^2} = \frac{n - n_{p0}}{L_n^2} \tag{2.22b}$$

where $L_p = \sqrt{D_p \tau}$ and $L_n = \sqrt{D_n \tau}$ are the hole and electron diffusion lengths, respectively.

For the current-voltage characteristic to be calculated, it is necessary to determine the diffusion components of the electron and hole currents flowing through the p-n junction and their sum

$$J = J_{p\,dif} + J_{n\,dif}$$

For this purpose it is necessary to know the densities of electrons and holes diffusing across the p-n junction. We start with the solution of the problem for the hole current, which then will be extended to cover the electron current. The hole density can be determined from equation (2.22a).

Since the minority-carrier density (holes in the given case) far from the p-n junction is equal to the equilibrium value, a boundary condition for the equation is defined by

$$p(x)|_{x=\infty} = p_{n0} \tag{2.23a}$$

Another boundary condition is determined by equation (2.18a) which describes the minority-carrier density at the p-n junction:

$$p(x)|_{x=0} = p_{n0}e^{qV/kT} \tag{2.23b}$$

The solution of equation (2.22a) will be given in the following form:

$$p = p_{n0} + Ae^{-x/L_p} + Be^{x/L_p} \tag{2.24}$$

Substituting equation (2.23a) into (2.24), we obtain $B = 0$. Considering that $x = 0$ for (2.24) and using (2.23b), we get

$$p_{n0}e^{qV/kT} = p_{n0} + A$$

whence it follows that

$$A = p_{n0}(e^{qV/kT} - 1)$$

The general solution of equation (2.24) will be thus written in the form

$$p(x) = p_{n0}(e^{qV/kT} - 1)e^{-x/L_p} \tag{2.25}$$

The hole diffusion current through the n-region

$$J_{p\,dif} = -qD_p\frac{\partial p}{\partial x} \tag{2.26}$$

Differentiating (2.25) with respect to x and substituting $\frac{\partial p}{\partial x}$ into (2.26) yield

$$J_{p\,dif} = \frac{qD_p p_{n0}}{L_p}(e^{qV/kT} - 1)e^{-x/L_p} \tag{2.27}$$

In a similar way one can obtain the expression for the electron diffusion current

$$J_{n\,dif} = \frac{qD_n n_{p0}}{L_n}(e^{qV/kT} - 1)e^{-x/L_n} \tag{2.28}$$

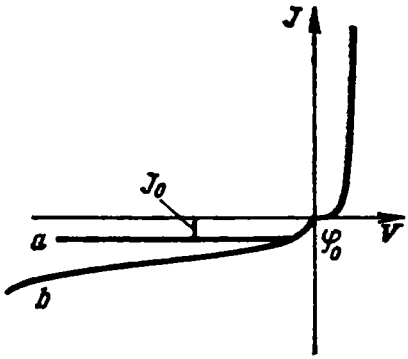


Fig. 2.8. Static current-voltage characteristic of a *p-n* junction (*a*) thermal current; (*b*) current due to generation of carriers in the *p-n* junction

For the plane $x = 0$ these expressions take the form

$$J_{n\,dif} = \frac{qD_n n_{p0}}{L_n}\left(e^{qV/kT} - 1\right) \tag{2.29}$$

$$J_{p\,dif} = \frac{qD_p p_{n0}}{L_p}\left(e^{qV/kT} - 1\right) \tag{2.30}$$

The total current through the *p-n* junction is equal to the sum of these currents:

$$J = q\left(\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right)\left(e^{qV/kT} - 1\right) \tag{2.31}$$

or

$$J = J_0\left(e^{qV/kT} - 1\right) \tag{2.32}$$

where

$$J_0 = q\left(\frac{D_p p_{n0}}{L_p} + \frac{D_n n_{p0}}{L_n}\right) \tag{2.33}$$

Expression (2.32) describes the current-voltage characteristic of the *p-n* junction.

In the forward-biased junction when $V \gg kT/q$, the current grows rapidly with increasing voltage. When $|-V| \gg kT/q$, the current through the *p-n* junction tends to reach a constant value J_0, called thermal current. As seen from (2.33), the higher the minority-carrier density, the larger the value of the thermal current, and thus, by (2.1), in order to diminish J_0 it is required to increase the density of majority carriers in the semiconductor.

The current-voltage characteristic of a *p-n* junction is shown in Fig. 2.8. In deriving equation (2.32) it has been assumed that the *p-n* junction is sufficiently thin and therefore the generation and recombination of carriers in the semiconductor may be disregarded.

The generation of carriers in the depletion layer gives rise to an additional reverse current

$$J_{0\,gen} = \frac{qn_i\delta}{2\tau} \tag{2.34}$$

This current is called generation current. Now we determine the ratio of generation current to thermal current.

Considering that the flows of electrons and holes are equal, the thermal current will be

$$J_{0\,dif} = \frac{2qn_{p0}D_n}{L_n} = \frac{2qn_{p0}L_n}{\tau}$$

The ratio of the currents is

$$\frac{J_{0\,gen}}{J_{0\,dif}} = \frac{n_i\delta}{4n_{p0}L_n} = \frac{n_n}{4n_i}\,\frac{\delta}{L_n} \tag{2.35}$$

Expression (2.35) reveals that for a semiconductor with a low lifetime of carriers and a small value of n_i the generation current in the space-charge layer exceeds the thermal diffusion current and therefore it must be taken account of.

Thus, in a silicon *p-n* junction the generation current exceeds the thermal current, and this is why its current-voltage characteristic has no saturation portion.

The current-voltage characteristic calculated from formula (2.32) and that plotted with regard for a generation effect, (a) and (b) respectively, are illustrated in Fig. 2.8.

2.5. EFFECT OF THE INJECTION LEVEL AND TEMPERATURE ON THE CURRENT-VOLTAGE CHARACTERISTIC OF A *p-n* JUNCTION

Equation (2.32) has been derived on the assumption that the current through the junction depends on the diffusion of carriers, which implies that the space charge of injected carriers is small and the electric field set up by this charge is insignificant. But this assumption is not always valid.

Let us introduce the notion of an injection ratio γ which is the ratio of the hole current to the total current through a *p-n* junction

$$\gamma = \frac{J_p}{J_p + J_n} \tag{2.36}$$

Substituting J_p and J_n from (2.29) and (2.30) into (2.36), then using the Einstein relation and replacing the minority-carrier density by the majority-carrier density, we obtain

$$\gamma = \left[1 + \frac{(\mu_n\mu_p)_n L_p\rho_p}{(\mu_n\mu_p)_p L_n\rho_n}\right]^{-1}$$

If we suppose that

$$\frac{(\mu_n\mu_p)_n L_p}{(\mu_n\mu_p)_p L_n} \simeq 1 \text{ and } \rho_p \ll \rho_n$$

then

$$\gamma = 1 - \frac{\rho_p}{\rho_n} \tag{2.37}$$

Substituting a typical value of $\rho_n = 5$ ohm·cm, and $\rho_p = 0.01$ ohm·cm gives $\gamma = 0.988$. This suggests that the current in an asymmetrical *p-n* junction depends on the carriers of one sign.

Now determine the injection level. Its quantitative characteristic is the ratio of the excess density of minority carriers to the equilibrium density of majority carriers.

Identifying the excess density of holes with $\Delta p_n = p_n - p_{n0}$ we obtain the following expression for the injection level:

$$\delta^* = \frac{\Delta p}{n_{n0}}$$

When $\delta^* \gg 1$, a noticeable gradient of the minority-carrier density appears in the vicinity of the *p-n* junction. This sets up an electric field and gives rise to a drift component along with a diffusion component of the current. The former component must not be ignored and therefore formula (2.32) does not hold any longer. The injected-carrier space-charge region offers an additional resistance to the electric current and part of the external voltage drops off across the space-charge region.

Calculations confirm that the injection level δ^* may be related to the forward current I_{for} flowing across the *p-n* junction by the following relation:

$$\delta^* = b\,\frac{I_{for}}{I_i}$$

where $I_i = kTS/q\rho_n W$; b is the ratio of mobilities; W is the thickness (width) of the *n*-region; S is the area of the *p-n* junction.

It is clear from the above that when the forward current I_{for} is bigger than I_i/b, expression (2.32) is inadequate. Let us note that in silicon with $\rho_n = 5$ ohm·cm, $W = 10^{-2}$ cm, $S = 10^{-2}$ cm^2, and $I_i = 5$ mA. If $b = 3$, expression (2.32) holds for current values less than 1.6 mA.

In deriving equation (2.32), the voltage drop across the semiconductor bulk resistance has not been taken into account. This bulk resistance causes redistribution of the applied voltage. Owing to the additional voltage drop across the bulk of the semiconductor and the injected-carrier space-charge region, the forward current-voltage curve becomes less steep.

This decrease in the slope of the curve can be mathematically described by introducing factor m into the exponent of the forward current-voltage characteristic.

Taking into account the injection level, the forward current-voltage characteristic for the p-n junction takes the form

$$I_{for} = \frac{2h}{\gamma} I_0 e^{qV/mkT} \tag{2.38}$$

where

$$h = \left(\frac{I_t}{bI_0}\right)^{\frac{m-1}{m}}; \quad m = 1 + \frac{W}{bL_p}$$

Let us examine the influence of the voltage drop across resistance r_r within the bulk of the semiconductor on the current voltage characteristic.

The voltage drop across the resistance r_r is given by

$$V_r = Ir_r$$

The voltage drop across the p-n junction is

$$V^* = V - Ir_r$$

Substituting this value into formula (2.32) we obtain

$$I = I_0 [e^{\frac{q(V - Ir_r)}{kT}} - 1]$$

or

$$V = \frac{kT}{q} \ln \left(\frac{I}{I_0} + 1\right) + Ir_r \tag{2.39}$$

So long as the current is small, the value $V_r = Ir_r$ is also small and the forward current obeys formula (2.32). But as the current increases the linear dependence of current on voltage V_r becomes more pronounced than the logarithmic dependence of current on voltage at the junction

$$\frac{kT}{q} \ln \left(\frac{I}{I_0} + 1\right)$$

As a result, when the current is heavy the current-voltage relation becomes linear and formula (2.39) can be replaced in this case by

$$I = \frac{V - V^*}{r_r} \tag{2.39a}$$

Figure 2.9 shows the forward current-voltage characteristic of a p-n junction plotted to a semilog scale with regard for the voltage drop across the injected-carrier space-charge region and the resistance within the bulk of the semiconductor.

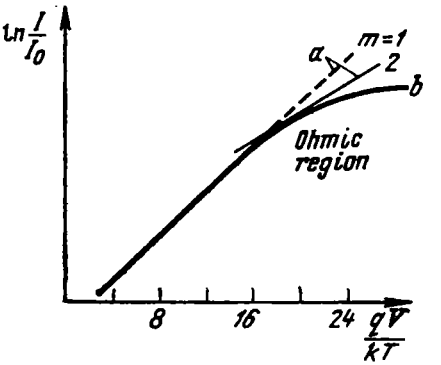


Fig. 2.9. Semilog plot of a *p-n* junction forward current-voltage characteristic
(*a*) design curve; (*b*) experimental curve

Now we consider the effect of temperature on the current-voltage characteristic of a *p-n* junction with a highly doped *p*-region. The current I_0 is generally determined in this case by the flow of holes across the junction, therefore

$$I = I_0 (e^{\frac{qV}{kT}} - 1) = \frac{qD_p p_n}{L_p} (e^{\frac{qV}{kT}} - 1)$$

We first examine the temperature dependence of a reverse thermal current through a *p-n* junction. Since $p_n = n_i^2/n_n$

$$I_0 = q \frac{D_p n_i^2}{L_p n_n}$$

In a semiconductor with fully ionized impurity atoms the temperature dependence of n_n may be disregarded. The temperature dependence of carrier density n_i in an intrinsic semiconductor has the form

$$n_i = N_c e^{-\frac{\Delta E}{2kT}}$$

where N_c is the density of states in the conduction band. Then

$$I_0 = q \frac{D_p N_c^2}{L_p N_d} e^{-\frac{\Delta E}{kT}} \tag{2.40}$$

or

$$I_0 = I_0 e^{-\frac{\Delta E}{kT}}$$

where I_0 is the quantity only slightly dependent on temperature. We presume the temperature dependence of ΔE, D_p, N_c and L_p may be neglected.

If the current I_0 at a certain temperature T_0 is known, then the current I_0 at another temperature T can be found by transforming

formula (2.40) into the form

$$I_0(T) = I_0(T_0)\, e^{a\Delta T} \tag{2.41}$$

For germanium and silicon the quantity a is 0.05 1/°C and 0.07-0.1 1/°C, respectively.

The current-voltage relation for the p-n junction biased in the forward direction is more complicated. At small values of the forward voltage, so far as $V < \Delta\varphi_d$ the temperature dependence of the current obeys expression (2.40). As soon as $V > \Delta\varphi_d$, the voltage drops off completely across the semiconductor bulk wherein all the impurity atoms are ionized, and the temperature dependence of the current is determined now by the temperature dependence of mobilities having the form $\mu \sim T^{-3/2}$ or $\mu \sim T^{3/2}$. At these voltages the forward current can diminish with rising temperature, in which case the forward and reverse currents through the p-n junction vary differently as the temperature rises, leading thereby to a decrease in the rectification ratio.

2.6. FIELD STRENGTH AND BREAKDOWN OF THE p-n JUNCTION

It has been shown in Sec. 2.2 that the width of the barrier in a p-n junction is related to the applied voltage by formula (2.16). If $V \gg \Delta\varphi_d$, then as follows from (2.16)

$$V = \frac{qN_d\delta^2}{2\varepsilon\varepsilon_0} \tag{2.42}$$

The mean field strength at a p-n junction may be calculated on dividing the voltage across the junction by its width

$$E = \frac{qN_d\delta}{2\varepsilon\varepsilon_0} \tag{2.42a}$$

or substituting the value of δ from (2.16)

$$E = \sqrt{\frac{qN_dV}{2\varepsilon\varepsilon_0}} \tag{2.43}$$

It is seen from (2.43) that with a rise in the reverse voltage the field strength at the p-n junction also increases, this increase being the greater, the higher the growth of the impurity density in the semiconductor.

When the field strength at the p-n junction reaches a certain critical value, its further growth is limited due to a sharp rise in the reverse current. This phenomenon is called breakdown of a p-n junction and the voltage at which the breakdown occurs is known as the breakdown voltage. The breakdown voltage can be expressed in terms of the critical field strength. For this purpose we transform (2.43) to the following form:

$$V = \frac{E_{cr}^2 2\varepsilon\varepsilon_0}{qN_d} \tag{2.44}$$

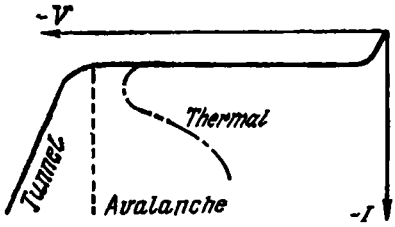


Fig. 2.10. Reverse current-voltage characteristic of a *p-n* junction under tunnel, avalanche and thermal breakdowns

There are three types of breakdown, namely, a field-effect (Zener or tunnel breakdown), avalanche (caused by collision ionization), and a thermal breakdown.

The first two types of breakdown are associated with increasing field strength at the *p-n* junction and the third, with a rise in power dissipation and therefore with growing temperature at the junction. The current-voltage characteristics of a reverse-biased *p-n* junction under various types of breakdown are presented in Fig. 2.10.

The nature of all the three types of breakdown is attributed to a sharp increase in the number of electrons flowing from the valence band to the conduction band of a semiconductor, which results in increasing density of electron-hole pairs.

This increase in the density of electron-hole pairs in case of Zener breakdown is due to a tunnel effect, i.e., tunneling of electrons over barriers the width of which is less than the wavelength of an electron. The barrier tunneling occurs without a change in the energy of an electron. The energy-band diagram of a tunneling junction and a breakdown process associated therewith is shown in Fig. 2.11.

The tunnel breakdown is commonly initiated at a field strength in the order of 2×10^5 V/cm for germanium and 1×10^6 V/cm for silicon. The probability of occurrence of a tunnel effect increases

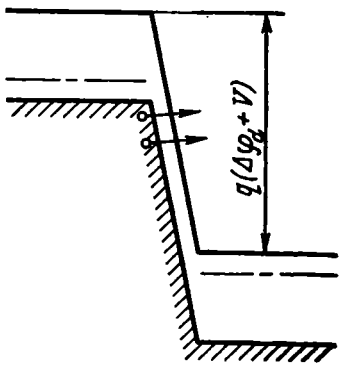


Fig. 2.11. Energy-band diagram of tunnel breakdown

with a decrease in the width of a *p-n* junction. Since the width of a *p-n* junction depends on the resistivity of a semiconductor (2.16), the tunnel breakdown voltage will rise with increasing semiconductor resistivity.

Indeed, substituting $\rho_n = 1/q\mu N_d$ into (2.44) yields

$$V_Z = 2\varepsilon\varepsilon_0\mu_n\rho_n E_{cr}^2 \tag{2.45}$$

The empirical equation for V_Z in germanium can be written as

$$V_Z = 99\rho_n + 48\rho_p \tag{2.46a}$$

and in silicon as

$$V_Z = 39\rho_n + 8\rho_p \tag{2.46b}$$

Compared to the tunnel breakdown, the avalanche breakdown occurs at a lower magnitude of the field strength. In the avalanche process the minority carriers acquire a sufficient energy at a *p-n* junction to effect impact ionization of a valence electron. The electron-hole pairs excited by the electron upon its collision with an atom lattice add to the flow of current through the *p-n* junction. These pairs are accelerated in turn by the electric field to an energy level sufficient to generate new electron-hole pairs, and the current through the junction grows avalanche-like.

For impact ionization to occur, an electron should have time to acquire a certain energy and this is why the avalanche breakdown is observed in those *p-n* junctions whose width is greater than the free path of an electron (i.e., when the resistivity of a base is fairly high).

In thin junctions (low-resistivity base), even high fields fail to adequately accelerate an electron to initiate impact ionization, and breakdown here is due to a tunnel effect.

Avalanche voltage V_a is related to resistivity ρ_r of a base by the empirical relation

$$V_a = a_a\rho_r^{m_a} \tag{2.47}$$

where a_a and m_a are characteristic numbers whose values are given in Table 2.1.

Table 2.1

Characteristic Numbers Used to Calculate Avalanche Breakdown Current through a *p-n* Junction

Material	Type of junction	m_a	a_a	n_a
Ge	p^+-n	0.6	83	3
	n^+-p	0.6	52	6
	p^+-n	0.65	86	5
Si	n^+-p	0.75	23	3

Since the breakdown current grows as a result of the avalanche multiplication of carriers, it may be written in the following form:

$$I = MI_0 \tag{2.48}$$

where I_0 is the reverse current when multiplication is absent, and M is the multiplication factor.

The breakdown of a *p-n* junction occurs at $M = \infty$. The empirical expression for the multiplication factor M is

$$M = \frac{1}{1-(V/V_a)^{n_a}} \tag{2.49}$$

The values of n_a for silicon and germanium are given in Table 2.1.

Substituting (2.49) into (2.48) we obtain the expression for the current-voltage characteristic of a *p-n* junction under the conditions of avalanche breakdown

$$I = \frac{I_0}{1-(V/V_a)^{n_a}} \tag{2.50}$$

As has been mentioned above, the breakdown process in a *p-n* junction depends on the resistivity of a semiconductor. At small values of ρ_r (less than 0.2 ohm·cm for germanium), a Zener breakdown is more likely to occur and at high values of ρ_r (more than 1 ohm·cm for germanium), an avalanche breakdown may take place. In the intermediate range of ρ_r values, both types of breakdown are equally probable.

Thermal breakdown takes place when the dissipated power heats the *p-n* junction. This heating causes an avalanche-like rise in current, i.e., breakdown of the *p-n* junction. Increasing the number of charge carriers due to heating may lead to a decrease in the voltage drop across the *p-n* junction, and the current-voltage curve of the *p-n* junction in the condition of thermal breakdown will have a portion with a negative differential resistance (Fig. 2.10).

The thermal breakdown voltage depends on the reverse current and is calculated from the formula

$$V_{th} = \frac{1}{I_0 R_{th} A} \tag{2.51}$$

R_{th} here is the thermal resistance between the junction and environment. This quantity is indicative of the amount of power required to raise the temperature of a junction by 1 °C.

For example, if $A = 0.09$ 1/°C, $I_0 = 20\mu$A, and $R_{th} = 0.5$ °C/mW, then $V_{th} = 400$ V.

In silicon junctions where current I_0 is small, as is clear from (2.51), thermal breakdown is practically nonexistent. The thermal breakdown voltage decreases appreciably with increasing temperature because in this case current I_0 begins to rise.

It should be noted that a more complicated type of breakdown is also possible: a rise in current due to a tunnel or an avalanche breakdown may result in a thermal breakdown, for at heavy currents voltage V_{th} turns out to be smaller than V_Z or V_a. This is why a vertical portion of the reverse current-voltage curve of a *p-n* junction (in case of a Zener or avalanche breakdown) may be followed by a portion of a negative differential resistance, which is typical of a thermal breakdown.

The breakdown develops more easily at the places where the *p-n* junction reaches the surface of a crystal. A semiconductor always has a surface charge which can appreciably distort the pattern of the field at a *p-n* junction by increasing or decreasing the width of the junction. If the surface charge coincides in sign with that of majority carriers in the base, the field strength at the junction near the surface of the semiconductor will be lower than within its bulk. The width of the *p-n* junction near the surface will thus be greater than in the bulk and the surface breakdown will not develop. But if the surface charge is opposite in sign to the majority carrier charge in the base, the *p-n* junction gets narrow near the surface and the probability of the surface breakdown increases.

2.7. *p-n* JUNCTION CAPACITANCE

The *p-n* junction capacitance is determined by a variation in the quantity of charge stored at the *p-n* junction due to a change in the magnitude of the external voltage applied to it.

Let us consider what defines the *p-n* junction capacitance. The total space charge in a *p-n* junction region per square centimetre of its area is given by

$$Q = \int_0^\delta qN_d\,dx = qN_d\delta \tag{2.52}$$

This charge, as is easily seen from (2.16), is dependent on the voltage applied to the *p-n* junction.

The quantity $C_{bar} = dQ/dV$, called the charge or barrier-layer capacitance, may be determined from expression (2.52) by substituting into it the value of δ from (2.16) and then differentiating it with respect to voltage:

$$C_{bar} = \sqrt{\frac{\varepsilon\varepsilon_0 qN_d}{2(\Delta\varphi_d + V)}} \tag{2.53}$$

The barrier capacitance of a *p-n* junction may be regarded as a capacitance of a parallel-plate capacitor with space δ between the plates. The higher the impurity concentration in a semiconductor, the larger the barrier capacitance because with a rise in N_d the *p-n* junction region becomes narrower.

For a graded p-n junction

$$C_{bar} = \sqrt{\frac{\varepsilon\varepsilon_0 q N'_d}{12(V + \Delta\varphi_d)}} \tag{2.54}$$

The p-n junction biased in the forward direction gets narrower and its barrier capacitance increases. But this capacitance is shunted by a small forward resistance across the p-n junction. A change in the value of the forward voltage across the p-n junction causes a change in the charge of the minority carriers injected into the n region. This effect can be regarded to be identical with that caused by a certain capacitance.

The minority-carrier density alters in the process of diffusion and therefore the capacitance associated with a variation of the charge in the n-region near the junction is called diffusion capacitance.

The density of the holes injected into the n-region is found from expression (2.25). When $e^{qV/kT} \gg 1$ expression (2.25) takes the form

$$p(x) = p_{n\,0} e^{\frac{qV}{kT}} e^{-\frac{x}{L_p}}$$

By integrating this equation, we can determine the total charge of minority carriers (holes) in the n-region per square centimetre of its area

$$Q_{dif} = q p_n e^{\frac{qV}{kT}} \int_0^\infty e^{-\frac{x}{L}} dx = q p_{n\,0} L_p e^{\frac{qV}{kT}} \tag{2.55}$$

The diffusion capacitance is

$$C_{dif} = \frac{dQ_{dif}}{dV}$$

Bearing in mind that the thermal hole current J_{0p} can be written in the form

$$J_{0\,p} = \frac{q p_n L_p}{\tau}$$

and that

$$J = J_0 (e^{\frac{qV}{kT}} - 1)$$

the differentiation of (2.55) yields

$$C = \frac{q}{kT} J_p \tau_p$$

where τ_p is the hole lifetime.

The total p-n junction capacitance may be regarded as a sum of the barrier and diffusion capacitances

$$C = C_{bar} + C_{dif}$$

In the forward-biased direction and at small reverse voltages, the diffusion capacitance is predominant and, conversely, in the reverse-biased direction and at $V \gg kT/q$ the barrier capacitance prevails.

The operation of *p-n* junction-type devices at high frequencies is affected by the *p-n* junction capacitance due to its bridging effect. Indeed, since the reverse capacitive reactance $R_C = 1/\omega C$, the rectification ratio of a *p-n* junction begins to decrease as the frequency rises.

2.8. OHMIC CONTACT

An ohmic contact is defined as a contact between a semiconductor and a metal, the current-voltage relationship of which is linear (ohmic). The ohmic contact is sometimes called nonrectifying contact. It must be low-resistive for the drop in voltage across it to be small.

The ohmic contact is distinguished for the absence of injection. Let us consider what happens if the contact on the *n*-side of a *p-n* junction becomes injecting. The holes injected by this contact will reach the *p-n* junction, increase its reverse current and thereby diminish the rectification ratio. This effect is particularly pronounced when the space between the ohmic contact and the *p-n* junction is comparable to the diffusion length of minority carriers.

It is practically impossible to make an ideal ohmic contact because barrier layers are always to be found at the boundary between a semiconductor and a metal, whereby the contact resistance often proves higher than the semiconductor resistance. To reduce the contact resistance a highly doped layer of low resistivity is made a short distance away from the contact (Fig. 2.12). The influence of injection, if any, by the ohmic contact is also diminished in this case since the lifetimes of minority carriers and their diffusion lengths shorten with an increase in the density of majority carriers.

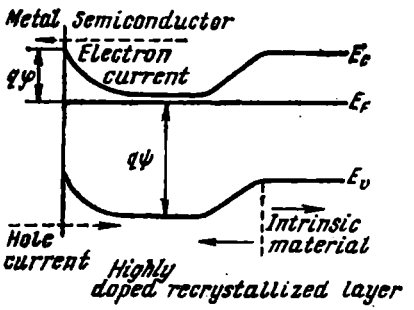


Fig. 2.12. Energy-band diagram of an ohmic contact between a metal and an *n*-type semiconductor

2.9. SURFACE STATE PROPERTIES OF A SEMICONDUCTOR

The conditions of atoms at the surface and in the bulk of a semiconductor are different. Indeed, in the bulk of a semiconductor, for instance, a silicon crystal, every atom having four valence electrons is surrounded by four neighbours so that each of the four covalent bonds of the atom is filled (Fig. 2.13*a*).

At the surface of the crystal every atom is linked only with three or fewer atoms and therefore each of them has at least one unsaturated bond. Such an unsaturated bond is able to capture a conduction electron and act as an acceptor, i. e., a negatively charged or a neutral centre.

The electron exchange reaction between an acceptor centre and the conduction band of a semiconductor has the form

$$Si^{(3)} + e^- \rightleftharpoons Si^{(3)-}$$

An atom with an unsaturated bond can also lose an electron, acting in this case as a positively charged or neutral donor centre. The electron exchange reaction of a donor centre with a conduction band of a semiconductor has the form

$$Si^{(3)} \rightleftharpoons Si^{(3)+} + e^-$$

The surface which has the states resulted only from the breaking of an ideal crystal lattice is referred to as an ideal surface. The density of states on the ideal surface is equal to the density of silicon atoms on the surface and totals 7.8×10^{14} cm^{-2}.

Actual surfaces have structural imperfections (for example, absence of an atom in the lattice site) and their surface state density can be higher than 7.8×10^{14} cm^{-2} (see Fig. 2.13*c*).

Besides, chemical and heat treatment used in the production of semiconductor devices can vary surface state densities over a wide range due in part to surface contamination, gas adsorption, etc.

The existence of surface states was first predicted by Tamm and experimentally confirmed by Shockley, Bardeen and Brattain.

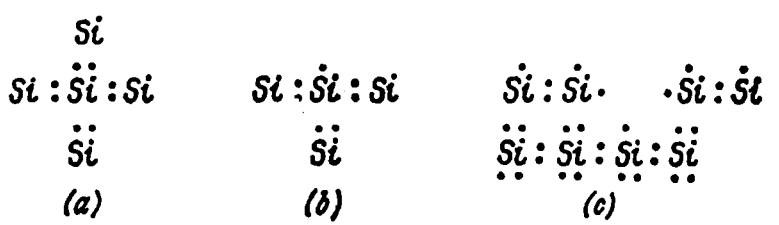


Fig. 2.13. Degree of saturation of covalent bonds in the bulk of an ideal silicon crystal (*a*), at the defect-free surface (*b*), and with one atom missing from the lattice site (*c*)

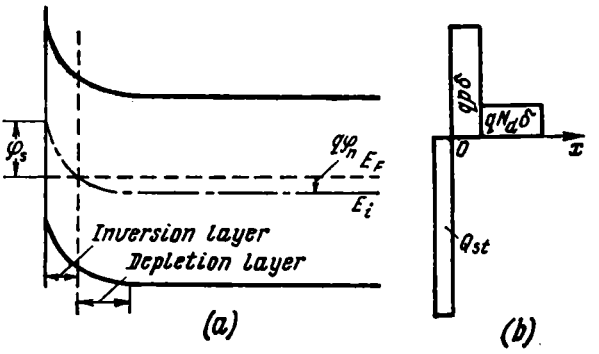


Fig. 2.14. Energy-band diagram illustrating acceptor states at the surface of an n-type semiconductor (a), and charge distribution on the surface (b)

From the standpoint of their energy levels, surface states are located in the forbidden band of a semiconductor near the bottom of the conduction band (donors) and the top of the valence band (acceptors).

By accepting or donating electrons, the surface states change the charge carrier density in the near-surface layer of a semiconductor, whereby, as has been described in Sec. 1.1, a space charge appears at the surface, and a potential difference $\Delta\varphi$ and an electric field develop between the surface and the bulk of the semiconductor.

Figure 2.14*a* shows an energy-band diagram of the surface of an n-type semiconductor containing acceptor surface states. The electron and hole densities, n_s and p_s, and n_b and p_b, respectively at the surface and in the bulk of the semiconductor are related to the electrostatic surface potential φ_s counted off from the Fermi potential by the relations

$$n_s = n_i e^{q\frac{\varphi_s}{kT}} = n_{b0} e^{q\frac{\varphi_s - \varphi_b}{kT}}$$

$$p_s = n_i e^{-q\frac{\varphi_s}{kT}} = p_{b0} e^{q\frac{\varphi_s - \varphi_b}{kT}}$$

The quantity $\Phi_s = \varphi_s - \varphi_b$ entering into these formulas corresponds to the bending of energy bands at the surface of a semiconductor. Depending on the type and amount of states on the surface of a semiconductor the following conditions hold.

1. The surface of a semiconductor is enriched in majority carriers (donor centres at the surface of an n-type semiconductor). An *accumulation* region of charge carriers is set up close to the surface. In this case φ_s and φ_b coincide in sign, $|\varphi_s| > |\varphi_b|$ and $\Phi_s < 0$.

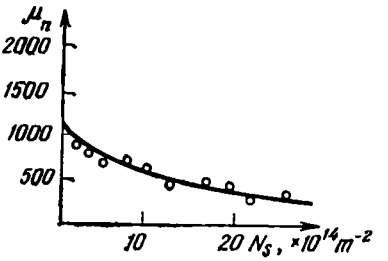


Fig. 2.15. Electron mobilities in an inversion layer at the surface of a *p*-type silicon crystal as a function of surface-state densities

2. The surface of a semiconductor is depleted of majority carriers (acceptor centres at the surface of an *n*-type semiconductor). In this case $|\varphi_s| < |\varphi_b|$ and φ_s and φ_b coincide in sign. A *depletion* region appears in the near-surface layer.

3. The type of conduction at the surface of a semiconductor is changed. A *p*-type layer develops at the surface of an *n*-type semiconductor. This effect is called *inversion*; it takes place when the number of surface states in a semiconductor is so large that the amount of conduction electrons in the space-charge layer proves insufficient to completely neutralize them. The unfilled surface states draw electrons from the valence band, forming holes in the near-surface layer, whereby φ_s and φ_b turn out to be opposite in sign.

Figure 2.14*a* shows that when the conduction of a semiconductor is inverted, it can also have a depletion region; Fig. 2.14*b* illustrates the distribution of a charge at the surface of a semiconductor. Surface states change not only the density but also the lifetime and mobility of charge carriers. Fig. 2.15 shows electron mobilities in the inversion layer at the surface of a *p*-type silicon as a function of surface state densities N_s.

Surface state properties have a pronounced effect on the parameters of semiconductor devices. Thus, a change in N_s may cause variations in the breakdown voltage of a *p-n* junction at the surface of a semiconductor, and a change in carrier lifetimes is responsible for changes in the parameters of photodiodes and phototransistors, and finally metal-insulator-semiconductor (MIS) transistors completely depend for their action on surface state properties.

Review Questions

1. How does the height of a potential barrier depend on the doping level of *n*- and *p*-regions of a *p-n* junction?
2. What is the height of a potential barrier at the boundary of a *p-n* junction in germanium whose n_n and p_p densities are equal to 10^{18} cm^{-3}?
3. Why do asymmetrical *p-n* junctions have extensive applications?
4. What determines the width of a space-charge region in a *p-n* junction? Find the width of this region in silicon with resistivities of 0.5 and 1 ohm·cm.

5. When are the formulas derived for an abrupt *p-n* junction applicable for describing a graded *p-n* junction?

6. What is the physical essence of diffusion length L?

7. When is the current through a *p-n* junction considered as a purely diffusion current?

8. What is the injection level and what effect does it have on the nature of motion of charge carriers across a *p-n* junction?

9. When and why does the forward current-voltage relationship of a *p-n* junction become linear?

10. Why does the avalanche breakdown whose critical field strength is lower than that of the tunnel breakdown develop at higher voltages?

11. Draw an energy band diagram for a *p-i-n* junction.

Chapter Three

DIODES

3.1. DIODE APPLICATIONS AND CLASSIFICATION

Among semiconductor devices, semiconductor diodes, i. e., two-terminal electric conversion elements, have found most extensive use.

Regarding their applications, diodes may be classified into the following groups.

1. Rectifier diodes intended to rectify alternating current of various frequencies and powers.
2. Pulse (switching) diodes intended for operation in pulse circuits.
3. Microwave detectors and switches. These diodes are used in microwave demodulator circuits (video detectors) and converter circuits. They must meet specific requirements for reactance and intrinsic noise levels, though their rectification ratios may be not too high. Gunn-effect diodes and avalanche transit time diodes also belong to this group.
4. Tunnel diodes designed to generate and amplify high frequency signals. These diodes feature a portion of negative differential conductivity on the forward current-voltage characteristic.
5. Voltage-variable capacitance diodes (varicaps) intended for use as elements with electrically controlled capacitance. They rely for their action on the dependence of the capacitance of a reverse-biased *p-n* junction on the voltage applied.
6. Avalanche diodes (Zener diodes) intended for stabilization (regulation) of voltage. Their action is due to a low dependence of voltage in the breakdown region on current.

By the type of *p-n* junctions one may differentiate (flat) junction diodes, point-contact diodes, point-junction diodes, and surface-barrier (junction) diodes.

In the flat *p-n* junction the linear dimensions determining its area are considerably larger than its thickness. In the point-contact junction all the linear dimensions determining its area are smaller than the width of the space-charge region. The point junction has nearly the same small area as the point-contact junction, the only difference being that the boundary between the *p*- and *n*-region in the point-junction is made flat. In the surface-barrier diode the *p-n* junction is produced due to the formation of an inversion layer at the surface of the semiconductor.

Low-frequency medium- and high-power semiconductor rectifiers (the cutoff frequency is not higher than 50 kHz) employ flat *p-n* junctions.

Point-contact and micro-alloy junction diodes serve to rectify low-power alternating current signals at frequencies up to 10^8 Hz.

This Chapter deals only with various-type rectifier diodes, while the diodes of special types — switching diodes, varicaps, Zener diodes, etc., — will be considered in the next Chapter.

3.2. METHODS OF PRODUCING *p-n* JUNCTIONS IN DIODES

Flat *p-n* junctions are made by introducing impurities into a semiconductor of a given conductivity type to form in it a region of the opposite conductivity type. A lower-resistivity side of the *p-n* junction, formed by alloying, is called the emitter and a higher-resistivity side is called the base.

By the method of preparing their emitters, junction diodes are divided into alloy-junction (fused-junction) and diffused-junction diodes.

An alloy junction is formed by fusing into a semiconductor a suitable metal or alloy which subsequently recrystallizes in it. The doping material contains acceptor or donor impurities depending on the conductivity type of the base. Thus, a germanium alloyed diode is made by fusing an indium-gallium alloy into an *n*-type germanium or an arsenic alloy into a *p*-type germanium crystal.

With the diffusion technique employed, impurities are introduced into a semiconductor from a gaseous medium by diffusion. For this purpose semiconductor wafers made of germanium or silicon are placed in a furnace filled with vapours of a doping material and are heated to the temperature close to the melting point of the semiconductor. The impurity atoms diffused into the semiconductor form on its surface a thin film (layer) of opposite conductivity type (*p*-type, for example); the penetration depth of the atoms forming the film depends on the temperature and duration of the diffusion process.

Since the diode requires the *p*-type layer only on one side of the crystal, the undesired portions of the film on its three remaining sides are removed by etching, and then the *p*- and *n*-regions are connected to ohmic contacts. This method enables one to prepare quality *p-n* junctions practically of any size, namely, from thousandths of a square millimetre to several square centimetres. It also allows an excellent control over the depth of penetration of a *p-n* junction and impurity density distribution in the crystal.

The law of variation of the impurity density in a semiconductor depends on the conditions of the diffusion process. The impurity diffusion is described by the Fick equation that has the form

$$\frac{\partial N}{\partial t} = D \frac{\partial^2 N}{\partial x^2} \tag{3.1}$$

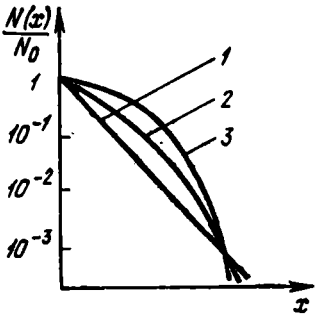


Fig. 3.1. Distribution of impurities in the base
1—calculated by the exponential law; 2—by formula (3.3); 3—by formula (3.5)

where $N(x, t)$ is the density of the diffused impurity; x is the coordinate counted off from the surface into the bulk of the material; t is the time, and D is the diffusion constant of the impurities.

When at the surface ($x = 0$) impurity density $N = N_0 = \text{const}$ at any moment of time t, and at $x = \infty$ the impurity density $N = 0$, the solution of the diffusion equation takes the form

$$N(x) = N_0 \left(1 - \frac{2}{\sqrt{\pi}} \int_0^{\lambda} e^{-\lambda^2} d\lambda\right) \tag{32.}$$

where

$$\lambda = \frac{x}{2\sqrt{Dt}}$$

or

$$N(x) = N_0 (1 - \operatorname{erf} \lambda) = N_0 \operatorname{erfc} \lambda \tag{3.3}$$

where $\operatorname{erf} \lambda \frac{2}{\sqrt{\pi}} \int_0^{\lambda} e^{-\lambda^2} d\lambda$ is the error function whose approximate values are tabulated.

Use can be made of the following approximation formula for many practical purposes:

$$N(x) = N_0 e^{-(1.08\lambda + 0.78\lambda^2)} \tag{3.4}$$

The surface density of impurities diffusing into a semiconductor from a gaseous phase depends on the solubility limit of the diffusate in the semiconductor and this density remains constant during the entire diffusion process. This is why the case considered here refers to the so-called diffusion from an infinite source.

The impurity density in a diffused p-n junction varies according to the law that approximates the exponential one (Fig. 3.1). The density gradient so produced gives rise to an electric field in a diode base, which impedes the motion of injected minority carriers. As

will be shown below in Chapter Four, the diodes with an electric field in the base region display good pulse characteristics. An example of these diodes are the КД503 type silicon diodes.

Apart from purely alloyed or diffused diodes, there are also alloyed-diffused diodes in which the junction is made in two stages — first by fusion and then diffusion to produce a density gradient in the base region. This method combines the low cost and simplicity of the fusion technique with the advantages of the diffusion technique (electric field in the diode base region).

In this method, impurities diffuse from a thin recrystallized layer. Since the number of atoms, per unit area, diffusing into the bulk of the semiconductor is limited, the diffusate density N_0 does not remain constant at the surface as it does in the case of diffusion from the gaseous phase. This type of diffusion is referred to as the diffusion from a limited source.

The solution of the diffusion equation for this case has the form

$$N(x) = \frac{Q}{2\sqrt{\pi Dt}} e^{-\frac{x^2}{4Dt}} \tag{3.5}$$

where $Q = N_0 a$ is the number of atoms per unit area of a thin layer with depth a.

When approximate calculations are made, the impurity distribution laws found from formulas (3.4) and (3.5) are often replaced by the exponential law $N(x) = N_0 e^{-x/2\sqrt{Dt}}$. The comparison made among the three laws is illustrated in Fig. 3.1.

In point (or point-contact) diodes the emitter is made in the form of a metal needle pressed against the surface of a semiconductor. Although the rectifying effect produced at this contact is rather good, the stability of contact characteristics is low and, besides, the rectification ratios vary from sample to sample.

To improve the quality and stability of the contact, use is made of an electric forming process based on sending high-power burst pulses across the contact in the forward or reverse direction. The power and duration of the pulse must be sufficient to bond the point of the metal needle with the semiconductor.

During the forming process the conductivity type of the semiconductor portion lying under the contact changes as a result of bonding. Thus, if a needle made of beryllium bronze is forced against an original plate of n-type germanium, then upon bonding the beryllium diffuses into the semiconductor and forms an abrupt p-n junction as it is an acceptor with respect to germanium. A p-n junction in the point diode is generally considered as a hemisphere (Fig. 3.2). It is just the small junction area of the point diode that determines a major distinction between the point and the junction diode.

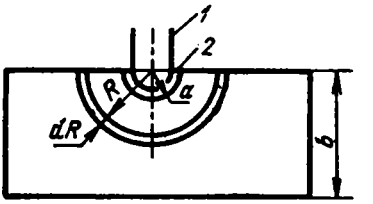


Fig. 3.2. Point-contact geometry adopted in calculating the spread resistance

1—metallic contact; 2—*p*-type layer in the *n*-type plate

As point diodes have a small junction capacitance, it is more preferable to employ them for operation in the high frequency range, including microwave frequencies.

Compared to junction diodes, point diodes have a higher base resistance and a narrower range of allowable currents and voltages on account of a small emitter area.

The base resistances in these diodes usually make up hundreds of ohms, while the allowable powers do not exceed 10 mW and forward currents lie within the range of 10 to 20 mA. The flow of high currents through such diodes leads to a loss of forming of the *p-n* junction and sharp impairment of its properties.

At the same working voltages the injection level in point diodes is higher, due to a small *p-n* junction area, than in junction diodes. This is why formula (2.32) ceases to be valid at lower currents. The reverse branch of the current-voltage characteristic of point diodes has some specific features.

First, a high current density and poor heat removal typical of the point diodes on account of their small junction areas cause heating of the junctions and lead to an increase in the reverse current. The saturation portion on the reverse curve is therefore weakly defined or absent at all.

Second, heating may result in a thermal breakdown of a point diode and emergence of a negative differential resistance portion on the reverse branch.

The merits of junction and point diodes are displayed in the so-called micro-junction (or micro-alloy junction) diodes.

In the micro-junction diodes, a needle, for instance, a gold-plated one, is kept with its flat point against a semiconductor surface. On applying a formation voltage to the contact, gold diffuses into the *n*-type germanium and forms a flat *p-n* junction (Fig. 3.3). The junction area in these diodes is only 2-3 times larger than that in point diodes (but hundreds of times less than in junction diodes), and for this reason their capacitances are small as they also are in point diodes. But this enlargement of the junction area allows for a certain increase in permissible forward currents and, besides, an improved transfer of heat from the surface of the micro-junction assures a better saturation of the reverse branch of the current-voltage characteristic. A typical example of the micro-alloy junction

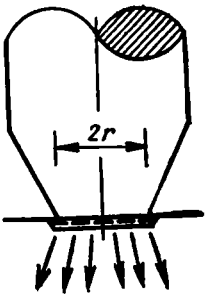


Fig. 3.3. Shape of the *p-n* junction in a micro-junction diode

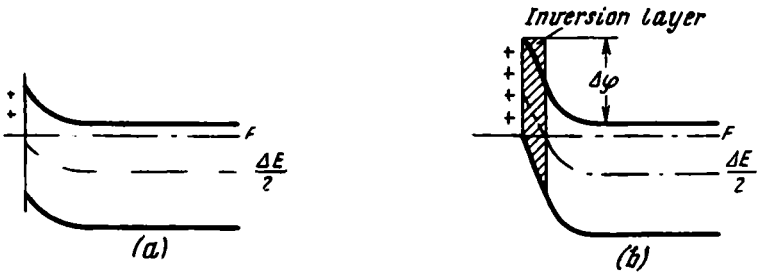


Fig. 3.4. Energy-band diagram of the contact between a semiconductor and a metal before inversion (*a*), and after the onset of inversion (*b*)

diode is the Д219A type silicon diode. In surface-barrier diodes the inversion layer is formed by electrochemically etching the semiconductor surface.

The inversion of conduction of a near-surface layer is due to the capture of conduction electrons by surface levels (Fig. 3.4).

3.3. EQUIVALENT CIRCUIT AND PARAMETERS OF DIODES

In the above Chapter we considered an ideal *p-n* junction, assuming that the leakage current in the *p-n* junction is small and therefore its reverse current depends only on the thermal current which is extremely low in junction diodes, especially in silicon diodes. In actual fact the leakage, mainly over the semiconductor surface, can sharply reduce the diode reverse resistance.

In analyzing the forward characteristic, account has to be taken of a semiconductor bulk resistance which is fairly large in point and micro-junction diodes. The current through the diode is governed in this case by resistance r_r of the semiconductor bulk and the applied voltage. This bulk resistance r_r is termed spread resistance

which can be found from the geometry of the contact illustrated in Fig. 3.2.

Let us denote the *p-n* junction radius by a and the crystal thickness by b. The resistance of a hemispherical layer of thickness dR at a distance R from the centre of the sphere may be written as

$$dr = \rho \frac{dR}{2\pi R^2}$$

The total resistance r_r is given by

$$r_r = \int_a^b dr = \frac{\rho}{2\pi} \int_a^b \frac{dR}{R^2} = \frac{\rho}{2\pi a} - \frac{\rho}{2\pi b}$$

Since $a \ll b$

$$r_r = \frac{\rho}{2\pi a}$$

With $\rho = 20$ ohm·cm and $a = 2 \times 10^{-5}$ cm, $r_r = 1.6 \times 10^3$ ohms. In practice, the value of r_r can be as low as tens of ohms due to the injection of minority carriers across the *p-n* junction.

Resistance r_r may be ignored when analyzing the reverse current-voltage characteristic. The backward impedance of the diode may be represented by a parallel combination of leakage resistance R_l and capacitance C_d.

Resistance R_l comes to hundreds and thousands of kiloohms in various diodes and C_d ranges from several units to tens of picofarads. The equivalent circuit of a diode is shown in Fig. 3.5.

Now we examine basic diode parameters which most completely describe the service properties of diodes (Table 3.1).

Table 3.1

Main Diode Parameters

Parameter	Unit of measurement	Parameter characteristic
I_{rev}	μA	Reverse current at a given value of reverse voltage
V_{rev}	V	Maximum allowable value of reverse voltage
V_{for}	V	Voltage drop across the diode at a given value of forward current through the diode
I_{for}	mA	Rectified current
C_d	pF	Diode capacitance at a given value of reverse voltage
Δf_c	kHz	Cutoff frequency up to which rectification is possible without reduction in rectified current
ΔT	°C	Working temperature range
η	%	Efficiency determined as the ratio of the product of rectified voltage and current to the active power consumed from the mains

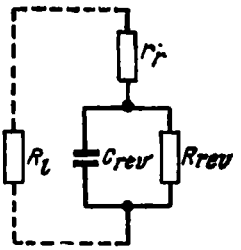


Fig. 3.5. Equivalent circuit of a diode

Forward voltage V_{for} relates to the so-called steady-state parameters of the diode. V_{for} is the voltage that drops off across a diode when forward current I_{for} of a certain value flows through it. The value of V_{for} in silicon and germanium rectifiers is 1.5 and 1 V, respectively.

Reverse current I_{rev} measured at some specified reverse voltage V_{rev} also belongs among steady-state parameters of diodes. Its value is usually determined at the maximum allowable reverse voltage $V_{rev\ max}$.

Steady-state parameters define the operation of diodes in direct-current circuits.

Dynamic parameters that characterize the performance of diodes in rectifiers include mean current I_0 and mean rectified forward voltage $V_{for\ m}$, both measured for one period at a given mean value of the forward current, mean reverse current $I_{rev\ m}$ measured for one period at a given value of the reverse voltage and mean reverse voltage $V_{rev\ m}$. The value of $I_{rev\ m}$ grows with a rise in the mean value of rectified current I_0 since heating of the diode by the forward current causes an increase in the reverse current.

The mean forward voltage depends on the diode forward current too, and therefore its value is specified at certain values of the forward current.

As is shown in the equivalent circuit (Fig. 3.5), the resistance of a reverse-biased diode is shunted by the *p-n* junction capacitance C_d, and thereby the backward impedance having an active and a reactive component diminishes with increasing frequency. The diode cutoff frequency Δf_{max} is the frequency at which the capacitive reactance becomes comparable in the order of magnitude to the *p-n* junction reverse resistance.

Therefore, for the upper frequency limit of diodes to be extended, the magnitude of capacitance should be brought to a minimum. It is due to the above reason that the values of capacitance C_d and cutoff frequency Δf_{max}, which also belong to the diode dynamic parameters, are specified. These parameters are usually determined by the *p-n* junction area; this is why the capacitance is the smallest in low-power point diodes and the largest in high-power junction

diodes. Accordingly, the upper frequency limit in point diodes exceeds by many orders of magnitude the working cutoff frequency of high-power diodes.

The properties of pulsed mode diodes are defined by the parameters associated with the effect of charge accumulation. These parameters will be described in the next Chapter.

To characterize maximum permissible operating conditions under which diodes are able to operate, the so-called maximum permissible parameters are introduced. Thus, the rectified voltage is limited by the *p-n* junction reverse breakdown voltage and for this reason it is required that maximum rated voltage $V_{rev\ max}$ be stipulated, which must be lower than the breakdown voltage.

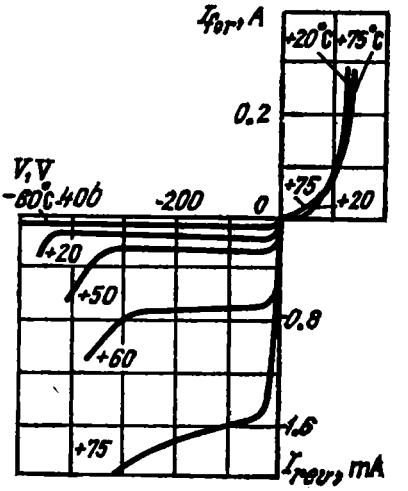


Fig. 3.6. Current-voltage characteristics of a diode at different ambient temperatures

The power dissipated by the diode biased in the forward direction at a high forward current may lead to impermissible heating of the *p-n* junction and increase in the reverse current. A maximum permissible value of forward current $I_{for\ max}$ is therefore specified for each type of diode.

It stands to reason that the better the removal of heat from a *p-n* junction and the larger its area, the higher the permissible current $I_{for\ max}$ and the diode dissipated power P_{max} which depends on $I_{for\ max}$.

The working temperature range of a diode is limited by its maximum permissible operating conditions. The lower limit of this range (—60 °C) is dictated by the mechanical strength of indium and contact alloys, and the upper limit, by the temperature dependence of semiconductor conductivity.

The maximum permissible temperature of a medium, t_m, surrounding a diode and that of the junction, t_j, and also the diode dissipated power P_{max} are closely interrelated. The higher the P_{max}, $I_{for\ max}$ and $t_{j\ max}$ associated with them, the lower the permissible temperature t_m. For this reason, lower- and medium-power diodes can operate at ambient temperatures of plus 75 °C (for Ge) and plus 125 °C (for Si), while high-power diodes require special cooling of junctions.

The dependence of forward current on temperature is somewhat more complicated. At small forward voltages when all the external voltage is applied to the *p-n* junction, the current grows with temperature in conformity with (2.40).

At high forward voltages $V > \Delta\varphi$ (0.3 V for Ge), all the voltage drops off across the bulk of a semiconductor and a change in current is determined by the temperature dependence of carrier mobility. Fig. 3.6 shows the current-voltage characteristics of a diode at different temperatures.

3.4. GERMANIUM AND SILICON RECTIFIER DIODES

Germanium and silicon junction rectifiers serve to rectify an alternating current of low and medium frequencies, usually up to 50 kHz.

Silicon diodes work at higher permissible working temperatures and hence higher permissible forward currents than germanium diodes due to a greater band gap width of the former. The resistivity of silicon diodes and the forward voltage drop across them are 1.5-2 times higher than in germanium diodes. This is due to a lower carrier mobility in silicon as against that in germanium. For the same reason, germanium diodes dissipate less power than silicon diodes at the same current.

As shown above (Sec. 2.5), the reverse current of a diode decreases with an increase in the width of a semiconductor band gap and thereby it is many times higher in germanium diodes than in silicon diodes. A wider band gap of silicon as compared with that of germanium ensures a higher maximum permissible temperature of silicon diodes, which is equal to 125-150°C.

The current-voltage characteristics of a silicon diode (type Д209) and a germanium diode (type Д7) are given for comparison in Fig. 3.7.

Compared to germanium diodes, silicon diodes withstand higher reverse voltages. The probability of thermal breakdown in silicon is small due to its wide band gap. The operation of silicon diodes is therefore stable in the near-breakdown region. They may be connected in series to rectify current of high voltages.

A particular feature of silicon diodes is an increase of the breakdown voltage with temperature, which is explained by the fact that breakdown in silicon depends on the avalanche multiplication process. As the temperature rises the heat scatter of mobile charge carriers increases and their free path diminishes, and for the carriers to acquire an energy sufficient to ionize atoms in the crystal lattice, it is necessary to raise the electric field strength. Such a mechanism is responsible for an increase in the breakdown voltage with rising temperature.

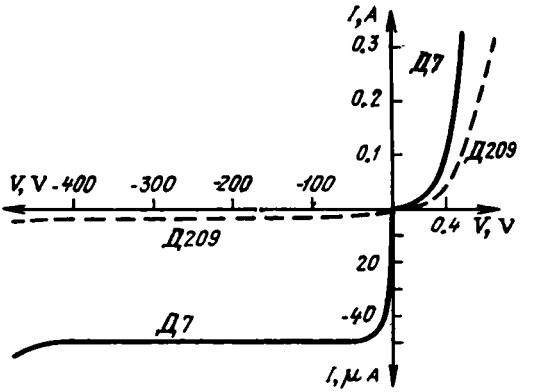


Fig. 3.7. Current-voltage characteristics of a junction diode and a point diode

This process must take place in germanium too; however, increasing temperature causes such a growth in the reverse current that the thermal breakdown whose voltage lowers with rising temperature begins to develop before the field strength has been increased.

Junction rectifiers are generally made by fusion. These diodes are capable of rectifying currents from 300 mA to 1000 A depending on the junction area and methods of cooling.

It has been mentioned in the previous Chapter that the properties of a *p-n* junction in the places where it reaches the surface are heavily dependent on the surface condition. Increased humidity, oxygen-containing medium and dust can appreciably change the carrier lifetime, recombination rate and *p-n* junction space-charge layer width at the surface, thereby causing variations in the parameters of semiconductor devices. To stabilize the surface conditions of a semiconductor and to protect the *p-n* junction area which reaches the surface against accidental contamination and mechanical damage, practically all semiconductor devices are hermetically sealed.

Diodes are enclosed in hermetically sealed ceramic or metal-glass cases. A special dry medium is maintained during the assembly of a device.

The design of low- and medium-power diodes is illustrated in Fig. 3.8.

In germanium diodes, the semiconductor crystal is soldered to the metallic crystal holder by the side opposite to the *p-n* junction. Soldering is done with a tin-antimony alloy providing an ohmic contact with germanium. In silicon diodes, the crystal is soldered to the metallic mount by the alloy which ensures an ohmic contact with silicon (gold and 1 per cent antimony for *n*-type silicon, and lead with 10 per cent admixtures of silver for *p*-type silicon). The

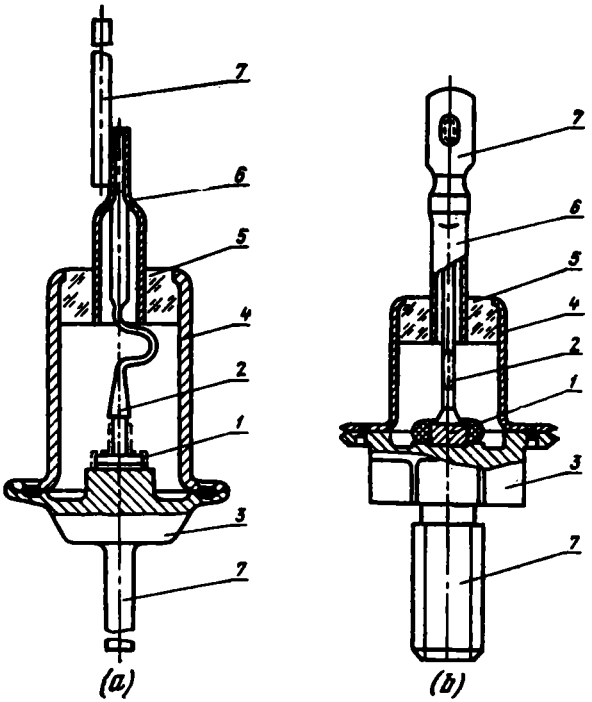


Fig. 3.8. Design of junction diodes of low power (*a*), and medium power (*b*)
1—crystal with a junction; *2*—inner lead; *3*—crystal mount; *4*—case; *5*—insulator; *6*—tube; *7*—external leads

power dissipated by a diode depends on the *p-n* junction area and case design. The *p-n* junction diameter in silicon diodes is chosen such that the permissible current density cannot exceed 200 A/cm^2.

The case for a low-power diode is made from kovar and is provided with a feed-through glass insulator moulded into it. Through this insulator extends a nickel lead connected with the *p-n* junction. Upon sealing the case around its periphery by pressure contact welding, the device is covered with black varnish to protect the crystal against penetration of light rays through the glass insulator.

The crystal mount in a medium-power diode also serves as a heat-sink and therefore it is usually fastened to the metallic chassis of a radio apparatus with screws (Fig. 3.8). In high-power diodes, a plate made of tungsten or kovar whose linear expansion coefficient is almost equal to that of silicon and germanium is interposed between the copper mount and the silicon or germanium crystal.

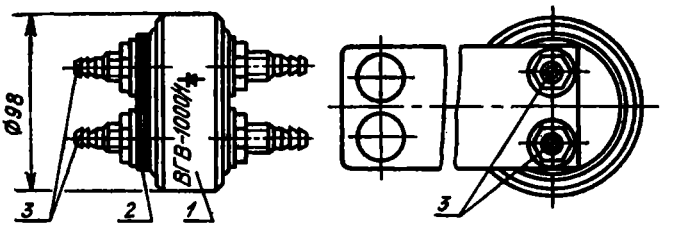


Fig. 3.9. Design of the ВГВ-1000 high-power germanium diode with water cooling
1—diode case; *2*—flexible output lead; *3*—water cooling outlets

Such an interlayer ensures against mechanical stresses arising due to some difference in these coefficients.

High-power diodes rated over 10 A require forced air or water cooling accomplished by special radiators.

The permissible values of forward current $I_{for\ max}$ and reverse voltage $V_{rev\ max}$ for diodes of small, medium and high power are listed in Table 3.2. The design of a ВГВ-1000 high-power germanium diode with water cooling is given in Fig. 3.9.

Table 3.2

Maximum Permissible Currents and Voltages for Some Types of Diodes

Type of diode	Cooling	$I_{for\ max}$, A	$V_{rev\ max}$, V
Low power, Д211	air	0.1	600
Medium power:			
Д214А	air	10	100
Д215	air	5	
Д305	air	10	50
High power:			
ВК-10	air	5	800
ВК-10	air, with radiator	10	700
ВК-10	air (15 m/s)	20	600
ВК-50	air	25	600
ВК-50	air (15 m/s)	100	600
ВК-200	air	50	600
ВК-200	air (15 m/s)	200	700
ВК-350	water (3 l/min)	350	700
ВГВ-1000	water (1 l/min)	1000	150

Germanium and silicon point diodes are practically similar in design. The semiconductor crystal is bonded to the mount (to provide an ohmic contact), and the *p-n* junction is produced by the

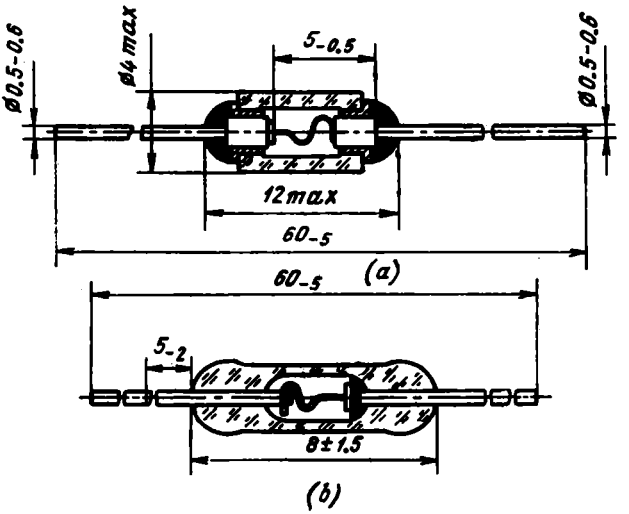


Fig. 3.10. Design of a point diode
(*a*) in a metal-glass case; (*b*) in a glass case

electric forming of a thin metal wire kept close against the crystal surface. The cases for point diodes can be made from metal-glass, glass or ceramic material.

The resistivity of silicon employed in the production of point diodes is 3 to 8 ohm·cm and that of germanium, 10 to 30 ohm·cm. The design of a point diode housed in a metal-glass case (*a*) and a glass case (*b*) is illustrated in Fig. 3.10.

The forward and reverse currents in Д104-type point diodes make up 30 and 0.01 mA, respectively, whereas the reverse voltage in the point diodes can reach 350 V. The barrier capacitance in the *p-n* junctions of these diodes is small and does not exceed 1 pF (0.7-0.6 pF for the Д104 diode).

The diodes of the above types rectify alternating currents in comparatively low-voltage circuits (up to 700 V). Higher voltages are rectified by series-connected low-voltage diodes. The diodes selected for this purpose should have approximately equal values of reverse resistances, otherwise all the reverse voltage may drop across the diode with the highest reverse resistance, causing its breakdown followed by the breakdown of the remaining series-connected diodes.

Semiconductor industry turns out rectifier stacks consisting of 4 to 14 thoroughly assorted silicon or germanium diodes. These stacks have the form of single units hermetically sealed in metallic or plastic cases. Some of the stacks (for instance, a double silicon rectifier

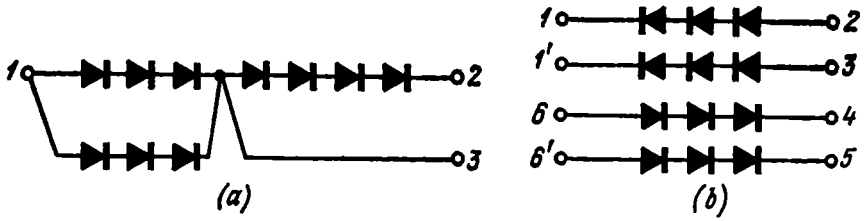


Fig. 3.11. Connection diagrams of the КЦ401А (*a*) and КЦ401Б (*b*) rectifier units

stack, type Д1009А) are connected into a circuit made up of two independent branches, each including five series-connected diodes. Rectifiers with still more complicated circuitry are also available. Thus, the circuit of silicon rectifier units, types КЦ401А and КЦ401Б, consists of three and four branches, respectively. The connection diagrams of these diodes assembled into units are shown in Fig. 3.11.

The maximum reverse voltage in each branch of these units reaches 500 V.

3.5. POLYCRYSTALLINE RECTIFIERS

Germanium and silicon rectifiers have found wide use only since the fifties of this century after the technique of growing single crystals of these substances was developed. Until then the alternating current of commercial frequencies was rectified by cuprous-oxide, titanium-dioxide and selenium polycrystalline rectifiers. They depend for their operation on the rectifying properties of the contact between a metal and an *n*-type semiconductor (titanium dioxide TiO_2) or *p*-type semiconductor (cuprous oxide Cu_2O and Se).

All these rectifiers are of the film type: a semiconductor and one or both contacts are made in the form of thin films.

Although the rectification ratios and efficiencies of these rectifiers are not high, the ease of manufacture, low cost and other beneficial features make them appreciably popular. The polycrystalline rectifier is made of a stack of rectifying plates, the number and the type of connection of which determine the operating currents and voltages of the rectifier. The properties of the rectifying plates depend on the type of semiconductor film and the method of its preparation. Particularly extensive use has been made of selenium rectifiers: for their production the industry turns out more than 600 types of plates of various size.

The production technique of oxide semiconductors (Cu_2O and TiO_2) used for the preparation of cuprous-oxide and titanium-dioxide rectifiers is practically the same. A cuprous-oxide rectifier

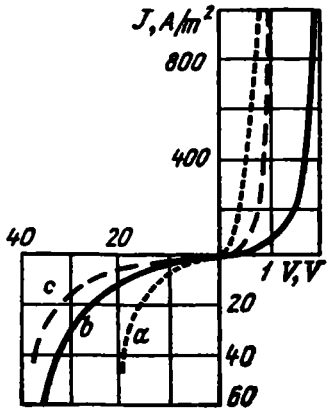


Fig. 3.12. Current-voltage characteristics of cuprous-oxide (*a*), titanium-dioxide (*b*), and selenium (*c*) rectifiers

is produced by oxidizing a copper plate at 1040 °C to obtain a cuprous-oxide layer over which a silver or gold film is deposited in vacuum. Rectification in such a two-electrode system takes place at the interface between copper and cuprous oxide, with the forward current flowing from the cuprous oxide to copper. The plate is then subjected to prolonged annealing at a temperature of 90°C to stabilize its parameters.

The current-voltage characteristic of such a plate is shown in Fig. 3.12*a*. It is seen from the drawing that the forward current-voltage curve has a linear portion. This feature of cuprous-oxide rectifiers is employed in measuring technique.

The maximum density of the current passing through a rectifying plate is 0.1 A/cm² and the maximum reverse voltage, 10 V.

The characteristics of cuprous rectifiers change with time due to ageing. Their properties are also heavily dependent on temperature. The temperature coefficients of reverse and forward currents, determined as relative changes in current per degree Celsius of temperature variation, are correspondingly $0.02°C^{-1}$ and $0.04°C^{-1}$. The working temperature interval of cuprous rectifiers stretches from minus 50 to plus 60°C.

The barrier layer in a titanium-dioxide rectifier is formed at the contact between an *n*-type titanium dioxide and gold or silver. The forward current corresponds to the flow of electrons from titanium dioxide to gold. The current density of a titanium-dioxide rectifying plate is 10-20 A/m² and the reverse voltage, 10 to 25 V. The current-voltage characteristic of such a plate is shown in Fig. 3.12*b*.

The rectifying ratios of titanium-dioxide rectifiers are not high, yet they can operate at a high working temperature up to 200°C and are capable of withstanding momentary heating to 400°C, which makes them suitable for a number of applications.

Selenium rectifying plates are fabricated on an aluminium or steel substrate covered with a thin layer of bismuth or nickel to form ohmic contact with a selenium film. The substrate so prepared is coated with a film of amorphous selenium by vacuum deposition, for example. Then the film is subjected to crystallization at a temperature close to the melting point of selenium, forming thereby

a *p*-type semiconductor layer which is subsequently plated with a cadmium-bismuth-tin alloy by metal pulverization (hot metal sputtering). Cadmium reacting with selenium forms a thin *n*-type layer of cadmium selenide (CdSe) and thus a *p-n* junction.

Rectifier parameters are then stabilized with the use of electric forming by reverse voltage, in which process cadmium diffuses into selenium and thus increases the thickness of the cadmium selenium layer.

Selenium rectifiers are highly reliable in service due to the fact that at the breakdown spot the overheated selenium first melts and then solidifies. The solidified selenium transforms into an amorphous state, i. e., becomes a dielectric, which results in "curing" of the breakdown spot. This phenomenon is used in the production of rectifiers to burn out "weak" spots by puncturing them with high-tension pulses.

The current-voltage characteristic of a selenium rectifier is illustrated in Fig. 3.12*c*. The electric parameters of selenium rectifiers approximate those of cuprous-oxide rectifiers. The efficiency of selenium rectifiers is not high, ranging from 45 to 60%. Notice that the efficiency of germanium and silicon diodes comes to nearly 100%.

The reverse voltage of selenium rectifiers ranges from 10 to 5000 V (type 5ГЕ200Ф) depending on the number of the selected plates, while the current, as has been stated above, is determined by the plate area. In the elements composed of plates 0.1×0.4 m² each, the currents run to 86 A (type 140ГМ24А6).

Ageing is typical of selenium rectifiers. Prolonged storage of selenium rectifiers may result in an increase of the reverse current by several times. This phenomenon is called loss of forming and is reversible: an alternating voltage applied to a long-stored rectifier restores the flow of a rated reverse current just in several minutes.

The loss of forming also takes place when the rectifier is kept energized at the forward voltage for a long time. This process is caused by the drift of cadmium ions into a space-charge layer and reduction in the resistance of the barrier layer.

Review Questions

1. What is the difference in the properties of alloyed and diffused *p-n* junctions and how does it influence the characteristics of diodes?
2. Why does the reverse current saturation not set up in point diodes?
3. What factors govern the forward voltage drop across a diode and which types of diodes have the highest and the lowest forward voltage drop?
4. What determines the maximum permissible temperature of a diode?
5. What are the design features of high-power diodes?
6. Calculate the spread resistance of a point diode formed in a germanium crystal with a resistivity of 5 ohm·cm and a *p-n* junction radius of 3×10^{-5} cm.
7. Why does breakdown in selenium rectifiers occur more rarely than in others?

Chapter Four

SPECIAL DIODES

4.1. SWITCHING DIODES

Semiconductor diodes have found extensive use in pulse circuits which serve a variety of purposes. High-speed circuits with a switching time of less than 1μs employ only specially designed diodes of fast response. These diodes are referred to as switching (pulse) diodes.

Let us consider transient processes determining the switching times of semiconductor diodes.

It has been stated in Chapter Two that under the applied forward bias holes are injected into a diode base and their density $p'_n = p_n e^{qV/kT}$ considerably exceeds the equilibrium density. The effect of emergence (or build-up) of an additional charge of injected minority carriers in the base is called storage or accumulation of carriers.

The injected holes diffuse into the bulk of the base, their density decreasing with the distance from the *p-n* junction due to recombination, and in a certain time the excess holes become distributed across the base as is shown in Fig. 2.7.

The time during which the holes get distributed throughout the length of the base depends on their lifetimes and determines the speed of transient processes in the diode biased in the forward direction.

Now consider the base resistance. At the first instant after the forward bias has been applied to the diode, the base resistance is governed by the steady-state conduction of the base material. This resistance is called pulse (switching) resistance.

The diffusion of the injected holes into the bulk of the base results in a decrease of the base resistance r_r which in the course of time is brought to a steady-state value determined by the value of p'_n. The time of the transient during which the forward resistance drops off to a steady-state value after fast biasing of the *p-n* junction in the forward direction is called the transient (recovery) time of the forward resistance.

Let us examine variations in the voltage across the diode on applying a square pulse. Diodes are usually connected so that the load resistance becomes by far greater than the forward resistance of the diode. The diode may thus be considered to be connected to a current generator. At the moment the forward current is thrown in the voltage drop across the diode is determined by the pulse resistance of the base and is equal to $V_{for\ max}$. After some time

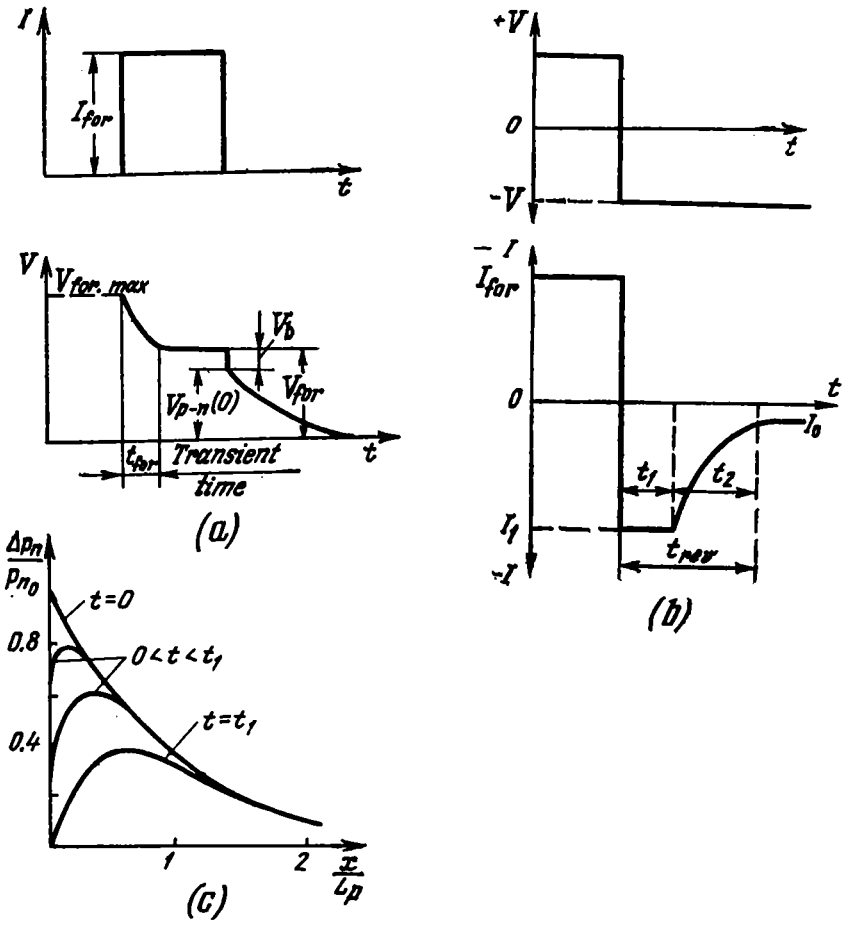


Fig. 4.1. Pulse shape of the forward voltage across the diode when connected to a current generator (*a*), current pulse shape on reversing the voltage across the diode (*b*), and distribution of holes across the base region on reversing the voltage across the diode (*c*)

t_{for} the diffusion of injected holes causes the base resistance to fall to a steady-state value and the voltage across the diode lowers to V_{for} (Fig. 4.1*a*).

For the junction diodes with a thin base ($W < 5L_p$), the time constant of this process is determined by the expression

$$\tau_{for} = 0.2(I_{for}\rho_r)^{-0.4}\, e^{W/L_p} \tag{4.1}$$

The time during which the voltage falls from $V_{for\ max}$ to 1.1 V_{for} is called the transient time t_{for} of the diode forward resistance or

the transient time of the forward voltage. Since τ_{for} varies for different types of diodes, the transient time ranges within (1-4) τ_{for}. The values of t_{for} and t_{rev} for some switching diodes are given in Table 4.1.

Table 4.1

Parameters of Switching Diodes

Type	Material of base and type of emitter	Recovery time of reverse resistance t_{rev}, μs	Transient time of forward resistance t_{for}, μs	Maximum forward voltage $V_{for\ max}$, V, at $I_{for}=50$ mA
Д18	germanium, point	0.07	0.08	1
Д20	germanium, point	0.1	0.08	5
Д219	silicon, micro-alloy	0.5	—	1
ГД507А	germanium, micro-alloy	0.1	—	0.5
1Д311	germanium, diffused	0.05	0.002	1.5
2Д503	silicon, diffused	0.01	0.002	—

As soon as the forward current is cut off the voltage across the diode surges up to a certain value of $V(0)$ and then gradually drops to zero. The value $V(0)$ is defined as the after-injection electromotive force (voltage).

Indeed, since the hole densities on both sides of the p-n junction are correspondingly equal to p_n' and p_{n0} due to injection, a potential difference $V_{p-n}(0)$ builds up across the junction, given by

$$V_{p-n}(0)=\frac{kT}{q}\ln\frac{p_n}{p_{n0}} \tag{4.2}$$

The surge in voltage after the forward current has been switched off is explained by the fact that at the moment of termination of the current pulse the ohmic drop in voltage across the base goes to zero, and since the p-n junction base layer retains the stored charge carriers, the after-injection voltage $V(0)$ still remains in the diode.

This voltage reduces due to the recombination of holes stored in the base and the process is described by the expression

$$V_{p-n}(t)=V(0)-\frac{kT}{q}\frac{t}{\tau_p} \tag{4.3}$$

when $kT/q \ll V$, and by the expression

$$V_{p-n}(t)=\frac{kT}{q}\left(e^{qV(0)/kT}-1\right)e^{-t/\tau_p} \tag{4.4}$$

when $kT/q \gg V$. Here τ_p is the lifetime of holes.

Now we analyze the transient process on reversing the sign of the applied voltage. Since the diode reverse resistance is commonly higher than the load resistance, the diode may be thought of as being connected to a voltage generator. The current pulse shape at reverse and forward voltages of equal magnitude and at a load resistance is illustrated in Fig. 4.1*b*. At the instant of the sign reversal the reverse current becomes equal in magnitude to the forward current and persists so for some time t_1 and then lowers within time interval t_2 to a steady-state value I_{rev}. The time taken for the reverse current to fall after the sign reversal to the value I_{rev} is called the recovery time t_{rev} of the reverse resistance. The reverse resistance recovery process is also associated with the storage of charge carriers in the diode base. Indeed, the steady-state value of the reverse current depends on the density of minority carriers in the base (see formula 2.32). Owing to the charge of holes stored in the base in the forward-bias state, the reverse current $I_{rev\,1}$ will by far exceed its steady-state value immediately after reversing the voltage.

The excess density of holes will reduce due to recombination and drawing of holes by the contact field from the base to the *p*-region of the diode. The distribution of holes across the base at different time intervals is shown in Fig. 4.1*c*.

The value of I_1 does not change until after the excess hole density at the *p-n* junction ($x = 0$) becomes equal to zero. This time t_1, i. e. the time of extinction of the base charge near the *p-n* junction, depends on the lifetime of holes and the values of forward (I_{for}) and reverse (I_1) currents, and in junction diodes it is related to these currents by the expressions

$$t_1 = 0.8\tau_p \left(\frac{I_{for}}{I_{for}+I_1}\right)^2 \quad \text{when} \quad \frac{I_1}{I_{for}} > 1 \tag{4.5}$$

$$t_1 = 0.65\tau \ln\frac{I_{for}}{I_1} \quad \text{when} \quad \frac{I_1}{I_{for}} < 0.2 \tag{4.6}$$

This time is less in point diodes; it depends on the diode hemisphere radius (see Ch. 3) and at $r_0 < 0.1\ L_p$ is determined by the relation

$$t_1 = \frac{r_0^2}{D_p}\left(\frac{I_{for}}{I_{for}+I_1}\right)^2 \tag{4.7}$$

When the density of the excess holes near the *p-n* junction becomes equal to zero the reverse current begins to fall. This process will continue till the density in all the regions of the base comes to equilibrium.

This decrease in the reverse current is described by the approximate expression

$$I_{rev} \simeq \frac{r_0}{L_p+r_0} I_{for} \frac{e^{-t/\tau_p}}{2\pi^{1/2}\left(\frac{t}{\tau_p}\right)^{3/2}} \tag{4.8}$$

The reverse-current fall time t_2 in junction diodes is (2-3) τ_p.

As with common diodes, the electrical properties of switching diodes are characterized by parameters V_{rev}, I_{rev}, I_{for} and C. Specific parameters of switching diodes are surge forward voltage $V_{for\ s}$ (at a given value of the current pulse), reverse-resistance recovery time t_{rev}, and forward-resistance transient time t_{for}.

The transient and recovery times determine the speed of response of the diode and one strives to reduce these values to a minimum. Thus, in diffused diodes the electric field of the base impedes the motion of injected holes and gives rise thereby to the storage of charge carriers in a thin layer near the *p-n* junction, with the result that upon reversing the voltage the time t_2 becomes far smaller than in alloyed diodes.

As has been stated above, quantities t_1, t_2 and t_{for} are related to hole lifetime τ_p. In order to reduce the storage time, the base is usually doped with impurities, such as gold, etc., which reduce the lifetime of holes. The diodes so produced have the reverse recovery times in the order of several nanoseconds. Besides, the forward recovery time is lowered by making the base as thin as possible. Examples of switching diodes are the ГD507A point, D219 microalloy and D311 diffused junction diodes.

The storing of charge carriers can be avoided if instead of a *p-n* junction a contact between a semiconductor and a metal is used as a rectifying element. The forward direction of current in such diodes accords with the flow of electrons from a semiconductor to metal and for this reason the charge carriers injected into the metal are cancelled practically instantaneously. The switching time of these diodes is less than 0.1 ns.

4.2. ZENER DIODES

The Zener diode, also termed avalanche or breakdown diode, is an electric solid-state device which exhibits on the current-voltage curve a portion with a low dependence of the voltage on the magnitude of the current flowing through it. This portion is observed on the reverse curve of the current-voltage characteristic of a silicon diode and is determined by the breakdown in the *p-n* junction (Fig. 4.2).

It is known from Sec. 2.6 that the silicon diode is mainly characterized by the tunnel or avalanche breakdown in the absence of negative resistance. Zener diodes are usually made from a high-conductivity material. The reverse current density up to the breakdown region is very small

$$j_0=\frac{n_i^2 D_n}{p_n L_n}=\frac{n_i^2 D_p}{n_n L_p}$$

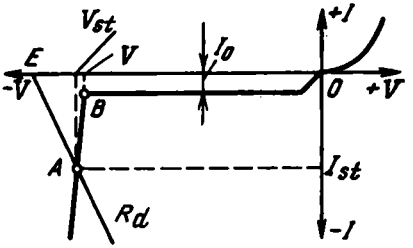


Fig. 4.2. Current-voltage characteristic of a Zener diode

and does not depend on voltage. This is why the reverse current-voltage curve may be approximated by straight lines diverging from the breakdown point at a right angle. The high steepness of the current-voltage curve enables one to adequately regulate the voltage.

The value of the regulation voltage (breakdown voltage) is determined by the width of a *p-n* junction. Hence, the higher the resistivity of the original silicon, the higher the regulation voltage. Low-voltage Zener diodes are manufactured from heavily alloyed silicon and their *p-n* junctions are made narrow, which causes the development of tunnel breakdown (see Sec. 2.6).

But if the *p-n* junction is made wide (more than the free path of charge carriers), avalanche breakdown tends to develop in it. The voltage at which the breakdown is likely to occur increases with growing resistivity of silicon. The required values of regulation voltages can thus be obtained by selecting the silicon of appropriate resistivity.

In principle, only tunnel breakdown occurs at regulation voltages below 6V, while at higher regulation voltages avalanche breakdown prevails. When it is required to stabilize low voltages of about one volt, use is made of the forward current-voltage characteristic at $V > \Delta\varphi_d$, which displays a weak dependence of voltage on current. The devices utilizing this effect are called stabistors.

The probability of thermal breakdown is negligibly weak in silicon diodes and therefore irreversible breakdown does not occur in Zener diodes.

Zener diodes are characterized by the following parameters:

1. Regulation (stabilization) voltage V_{st}, i. e., a voltage drop across a Zener diode at a certain regulation current flowing through it.

2. Regulation voltage spread ΔV_{st}, i. e., an allowable deviation of the regulation voltage from the value V_{st} at a given regulation current.

3. Regulation (stabilization) current I_{st} flowing through a Zener diode within a regulation (stabilization) region.

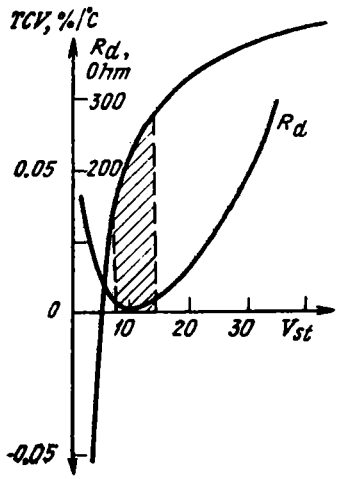


Fig. 4.3. TCV and differential resistance as functions of the regulation voltage

4. Differential resistance $R_d = \frac{dV}{dI}$ which is the ratio of the incremental voltage across a Zener diode to a small incremental current that caused the voltage increment.

5. Static resistance $R_{st} = \frac{V}{I}$ which is the ratio of the voltage across a Zener diode to the current through it.

6. Quality factor

$$Q_{st} = \frac{R_d}{R_{rev}}$$

7. Temperature coefficient of regulation voltage, TCV, which is the ratio of a relative change in the regulation voltage, expressed in per cent, to an absolute change in ambient temperature.

Since the working region of the Zener diode current-voltage curve has a certain slope, the regulation voltage depends on the choice of the operating point (point A in Fig. 4.2). In an ideal diode, this voltage coincides with the avalanche or tunnel breakdown voltage of the p-n junction.

In a real diode, as is shown in Fig. 4.2, these two voltages do not coincide; therefore, one should specify a certain regulation current I_{st} in order to determine an unambiguous value of the regulation voltage.

The slope of the current-voltage characteristic at the operating point (point A) determines a dynamic or differential resistance R_d of a Zener diode. The higher this value, the larger the variations in the voltage across the diode with a change in the regulation current, and therefore in designing Zener diodes the ways are sought for to obtain the lowest values of R_d.

The value of R_d is determined by the mechanism of breakdown. For the tunnel breakdown

$$R_d = \frac{BV_{st}}{I} \tag{4.9}$$

where B is the coefficient accounting for the properties of the material chosen; V_{st} is the breakdown voltage equal for silicon to

$$V_{st} = 39\rho_n + 8\rho_p$$

Under avalanche breakdown

$$R_d = \frac{V_a}{nMI} \tag{4.10}$$

where V_a is the avalanche breakdown voltage (see Sec. 2.6).

As is seen from formula (4.10), this resistance diminishes with increasing current through the *p-n* junction.

The effectiveness of a Zener diode is characterized by quality factor Q_Z. The steeper the diode characteristic curve, the smaller the quality factor value and hence the better the diode. In modern Zener diodes the values of the quality factor lie within 0.01-0.05 and even below.

The regulation voltage is related to the value of TCV. Fig. 4.3 shows the dependence of dynamic resistance R_d and TCV on the regulation voltage for different Zener diodes. It is clear that the TCV grows with increasing Zener diode voltage and the R_d has a minimum in the regulation region of V_{st} of about 7 V. This is due to the fact that at V_{st} reaching 5-7 V the tunnel and avalanche breakdowns develop simultaneously in the *p-n* junction. As the V_{st} increases the effect of tunnel breakdown diminishes and at voltages over 7-10 V the avalanche breakdown alone plays a dominant role.

It is the joint action of the two types of breakdown that is responsible for the minimum of R_d in the tunnel-to-avalanche breakdown transition region.

Figure 4.3 shows that though the values of TCV remain extremely low at all regulation voltages (not more than 0.1%/°C) yet they grow with voltage from 0.05% at 5 V to 0.1% at 50 V.

The shaded region from 7 to 14 V on Fig. 4.3 corresponds to the region of operating voltages of Zener diodes, types Д808 through Д813.

The TCV is positive in Zener diodes with $V_{st} > 5$ V, which is attributed to the fact that under the avalanche breakdown conditions the temperature dependence of the breakdown voltage V_a is determined by a reduction in the free path of carriers due to an increase in the number of collision of carriers with the lattice. The breakdown voltage must grow in this case (see Sec. 2.6).

In low-voltage Zener diodes noted for avalanche breakdown, the TCV is negative. This is explained by the fact that the voltage at which an electron tunnels from the valence band to the conduction band is associated with the width of a semiconductor band gap. The smaller the band gap width, the lower the avalanche breakdown voltage. And as its width diminishes with rising temperature, the breakdown voltage also decreases.

In order to reduce the value of TCV, temperature-compensated voltage regulators are made in which a reverse-biased *p-n* junc-

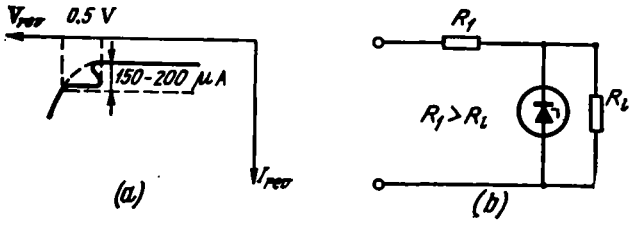


Fig. 4.4. Negative resistance region on a Zener diode current-voltage characteristic (*a*), and Zener diode connection diagram (*b*)

tion is connected in series with one or more forward-biased *p-n* junctions.

As is shown in Sec. 2.5, the forward voltage across the *p-n* junction diminishes with increasing temperature, i.e., TCV < 0. The TCV of such a temperature-compensated Zener diode will be equal to the algebraic sum of TCV of all the junctions and may be kept appreciably small.

Thus, in the type Д818 precision Zener diodes the main *p-n* junction is series-connected with two forward-biased compensating *p-n* junctions. In this manner it is possible to produce Zener diodes with a TCV not higher than 0.001%/°C.

The maximum regulation current is limited by the allowable heating of the *p-n* junction. However, a Zener diode cannot operate at particularly small currents, for in this case the breakdown voltage becomes dependent on the current as the number of carriers is too insignificant to initiate the avalanche breakdown. This is why maximum ($I_{st\ max}$) and minimum ($I_{st\ min}$) regulation currents are specified. Characteristic of Zener diodes (particularly high-tension diodes) is the generation of much noise in a transition range from the reverse-current saturation region to the breakdown region, the transition range being called the noise bandwidth of breakdown. This phenomenon is typical of avalanche breakdown. The noise bandwidth lies in the range of 0.2 to 0.3 mA; in this region the current-voltage characteristic has a negative resistance portion (Fig. 4.4*a*). The simplest connection diagram of a Zener diode is given in Fig. 4.4*b*. As the voltage rises, the resistance of the Zener diode diminishes and all the excess voltage falls off across resistance R_1, while the voltage drop across load resistance R_l remains invariable.

The parameters of some types of commercial Zener diodes and also the 2C107A silicon stabistor are listed in Table 4.2.

General-purpose Zener diodes, types Д808 through 813, КС156А, КС168А, etc., used in the circuits of voltage regulators and direct

Table 4.2

Parameters of Zener Diodes

Type	V_{st}, V	R_d, ohms (at 5 mA)	TVC, %/°C
2C107A	0.63	7	—2
2C139A	3.5	60	—0.1
2C168A	6.0	46	0.06
Д810	9-10.5	12	0.55
Д809	8.5-9.5	10	0.62
Д813	11.5-14	18	0.8
Д818	9	25	0.02
Д816	22	120	—
Д817	56	200	0.14
КС630А	130	1000	0.2

current limiters are made by fused-junction technique, and precision Zener diodes (for example, Д818 type) employed as standard voltage sources or high-precision reference elements are made by diffusion technique. As with other diode types, Zener diodes are sealed in metal-glass cases.

4.3. VARICAPS

It has been shown in Chaper Two that the barrier capacitance of a *p-n* junction is voltage dependent and so practically any semiconductor diode may serve as a voltage-variable capacitor.

Semiconductor diodes used as nonlinear voltage-depended capacitances and having low losses in the operating frequency bands are called varicaps (voltage-variable capacitance diodes). A varicap is convenient in that the constant bias impressed on it enables a remote and practically inertialess variation of its capacitance and thereby variation of the resonance frequency of the circuit that incorporates the varicap.

Varicaps are used for the amplification of microwave pulses, and also for the operation in frequency-tuning tanks and automatic control systems. The barrier capacitance of a *p-n* junction is essentially frequency independent, and as its quality factor is sufficiently good and noise level low, the advantage of employing varicaps in microwave frequency circuits is evident.

As has been shown above in Chapter Two, the forward-biased diode also has a diffusion capacitance, its value being even higher than that of the barrier capacitance. The diffusion capacitance depends on voltage, as does the barrier capacitance, but the poor quality factor and high noise level of this capacitance associated with fluctuations in the injection current over the junction area make it unsuitable in circuits.

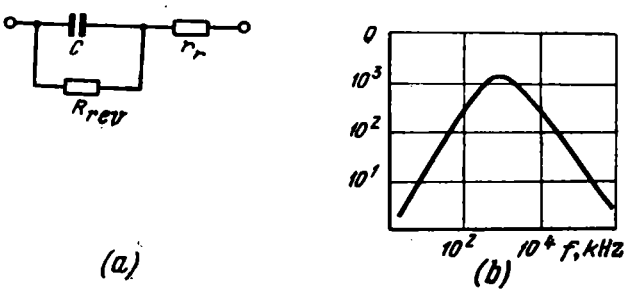


Fig. 4.5. Equivalent circuit for a varicap (*a*) and its quality factor-frequency relationship (*b*)

By changing the voltage across a varicap, it is possible to vary its capacitance over a wide range. The range of this voltage variation is characterized by the capacitance ratio which is the ratio of the maximum to minimum capacitance.

The operation of the varicap as an element of a circuit may be illustrated by means of an equivalent circuit (Fig. 4.5*a*) in which the *p-n* junction is represented by capacitance C shunted by reverse resistance R_{rev}. Connected in series with the junction are resistance r_r across the semiconductor bulk and inductance L_l of output leads. In addition, the circuit is shunted by capacitance C_c of the varicap case, which comes to 1-1.5 pF and may be disregarded at low frequencies; the same refers to the output lead inductance L_l ranging within several nanohenries. This is why the case capacitance C_c and lead inductance L_l are not indicated on the circuit diagram.

The resistance r_r acquires a particular importance at higher frequencies. Indeed, the impedance z of the circuit (Fig. 4.5) is equal to

$$z = r_r + \frac{\frac{1}{j\omega C} R_{rev}}{R_{rev} + \frac{1}{j\omega C}} \tag{4.11}$$

where C is the diode capacitance determined from formula (2.53).

Reducing the expression to the common denominator and multiplying the numerator and denominator by $1 - j\omega C R_{rev}$ yield

$$z = \frac{r_r \omega^2 C^2 R_{rev}^2 + r_r + R_{rev} - j\omega C R_{rev}^2}{\omega^2 C^2 R_{rev}^2 + 1} \tag{4.12}$$

Hence, the capacitive reactance

$$X_C = \frac{\omega C R_{rev}^2}{\omega^2 C^2 R_{rev}^2 + 1} \tag{4.13}$$

Considering that in formula (4.12) $\frac{r_r}{R_{p\text{-}n}} \ll 1$, the expression for the quality factor may be reduced to the form

$$Q = \frac{X_C}{z} = \frac{\omega C R_{rev}}{\omega C^2 R_{rev} r_r + \frac{r_r}{R_{rev}} + 1} \tag{4.14}$$

The dependence of the varicap quality factor on frequency is displayed graphically in Fig. 4.5*b*. At low frequencies, when $r_r \ll \ll 1/\omega C$ and $r_r/R_{rev} < 1$, all the terms in the denominator of formula (4.14) may be disregarded as compared to unity and then

$$Q = \omega C R_{rev} \tag{4.14a}$$

At high frequencies the term $\omega^2 C^2 R_{rev} r_r \gg \frac{r_r}{R_{rev}} + 1$ and so the last two terms in the denominator may be neglected. Hence,

$$Q = \frac{1}{\omega C r_r} \tag{4.14b}$$

It is obvious that the higher the resistance r_r, the lower the value of the quality factor.

Formula (4.14b) shows that a quality varicap must be made from a material with a low value of ρ. But the resistivity can be decreased only to a definite limit since a reduction in the resistivity entails a narrowing of the *p-n* junction and capacitance variation range (decrease in the breakdown voltage). The value of r_r is therefore lowered by making the varicap base as thin as possible.

A varicap is prepared from a low-resistivity silicon plate covered with a thin high-resistivity layer of the same conductivity type, i.e., an n^+-n structure.

The p^+-n-n^+ structure is made by diffusing an acceptor impurity in the *n*-layer. The thin *n*-region in the structure serves as a base, while the resistance of the n^+-layer may be disregarded.

Varicaps are similar to common diodes in design. The electrical parameters of silicon varicaps are given in Table 4.3. The design of varicaps provides for the dissipation of power up to 250 mW and ensures stable operation at temperatures up to 140°C.

Table 4.3

Parameters of Varicaps

Type	Capacitance at 4 V, pF	Maximum bias, V	TCC at 4 V, 1/°C	Quality factor at 50 MHz	Capacitance ratio at 50 MHz
Д901A	22-32	80	5×10^{-4}	25	4
Д902	6-12	25	5×10^{-4}	30	3

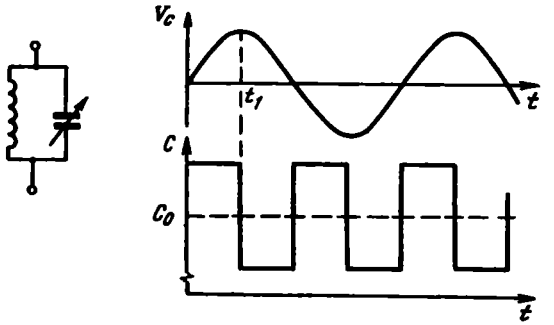


Fig. 4.6. Capacitance and voltage variations in an oscillatory circuit during parametric amplification

A metal-insulator-semiconductor capacitor (MIS capacitor) can also serve as a voltage-variable capacitor.

If the voltage is applied to a MIS system so that the n-type semiconductor is connected to a positive side, then a layer depleted of charge carriers will be formed on the semiconductor surface under the insulator. Since this layer may be considered as a space charge layer, the total capacitance of the MIS capacitor may be written in the form

$$C = \frac{C_i C_d}{C_d + C_i} \tag{4.15}$$

where C_d is the capacitance dependent on the depletion layer, and C_i is the capacitance of the insulator.

By selecting a semiconductor resistitivity it is possible to achieve manifold variations in the capacitance. However, the presence of surface states and exchange of electrons between the states and the conduction band lead to the formation of one more capacitance highly dependent on frequency. For this reason, in designing varicaps, use is more commonly made of the barrier capacitance which is little dependent on frequency.

Owing to their high quality factor, varicaps can be used to amplify and generate oscillations in parametric amplifiers. Examine the operation of a parametric amplifier.

We first couple a variable capacitor to a parallel circuit and set up oscillations therein. Now reduce the capacitance of the capacitor (widen the p-n junction by the constant voltage applied) at the moment t_1 when the voltage across the capacitor is maximum (Fig. 4.6). As the voltage and the charge stored by the capacitor are related by the expression $C = dQ/dV$, the invariable charge on the capacitor will cause an instantaneous increase in the voltage across it, in other words, the energy expended on reducing the capacitance will be converted to the energy of an electric field in the capacitor.

If we then bring the capacitance of the capacitor to its original value, the voltage in the circuit will decrease for the same reason and the circuit will return the energy consumed. If the capacitance is brought to its original value at the moment when the voltage across the capacitor goes through zero (there is no charge on the plates), such a change will not cause the transfer of the energy from the oscillatory circuit to the variable capacitor. The energy expended on varying the capacitance is transformed into the energy of an oscillatory circuit.

Thus, decreasing the capacitance of the capacitor at the moment the voltage reaches its peak and increasing it at the moment the voltage goes through zero, enable' us to pump energy into the oscillatory circuit and amplify oscillations.

It is evident from Fig. 4.6 that the rate of variations in the capacitance must be twice as high as the rate of the oscillations being amplified.

The control voltage effecting variations in the varicap capacitance is also called pump voltage.

As purely reactive elements do not produce intrinsic noise, the noise level of parametric amplifiers is extremely low.

The working frequency of varicaps is limited by the dielectric relaxation time τ (the drift time of majority carriers in the p- and n-regions under the influence of an electric field) which is equal to

$$\tau = \varepsilon\varepsilon_0\rho \tag{4.16}$$

Setting $\varepsilon = 16$ and $\rho = 0.1$ ohm·cm, the value of τ for germanium will equal 1.4×10^{-13} s.

The working frequency of real varicaps is restricted by the losses due to the junction forward resistance r_r (see Fig. 4.5).

If the minimum quality factor of the oscillatory circuit is assumed to be equal to

$$Q = \frac{1}{2\pi f C r_r} = 1 \tag{4.17}$$

then, allowing for

$$r_r = \frac{W}{q\mu n S} \tag{4.18}$$

where W is the width of the base, and substituting the value of C_{rev} from (4.14) and of r_r from (4.18) into formula (4.17), we find the expression for the cutoff frequency

$$f = \sqrt{(V + \Delta\varphi_d)\frac{q\mu^2 n}{2\varepsilon\varepsilon_0}}\,\frac{1}{W\pi} \tag{4.19}$$

It is evident from (4.19) that the higher the density and mobility of carr i ers in the n-base, the greater the value of f.

The operating frequencies of parametric amplifiers now extend to microwave ranges. Thus, the parametric amplifiers rated at a frequency of 16 GHz with an amplification of 20 db have a noise factor of 2.2 db at a bandwidth of 900 MHz. It has been reported recently that under development are parametric amplifiers embracing a frequency range of more than 90 GHz.

4.4. TUNNEL DIODES

Tunnel diodes belong to the devices exhibiting a negative differential resistance which enables the tunnel diodes to amplify and generate signals.

The tunnel diodes display the so-called N-shaped forward curve on the current-voltage characteristic (Fig. 4.7).

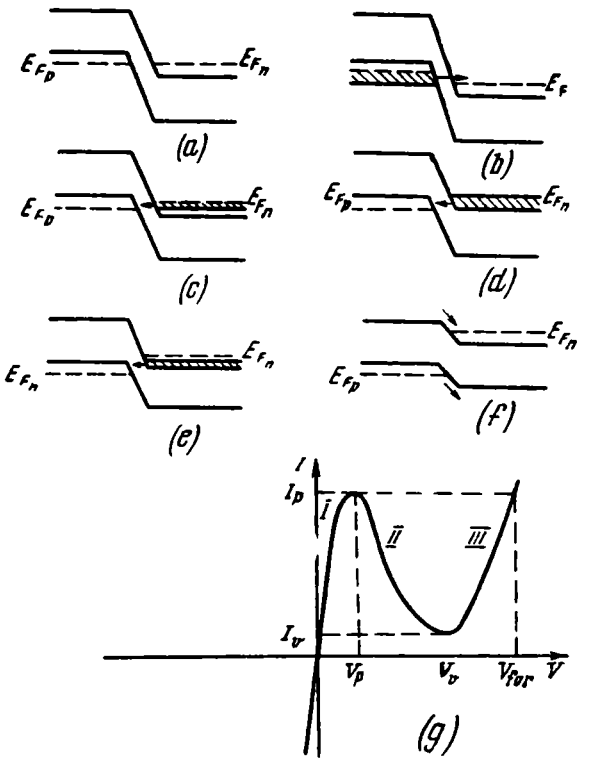


Fig. 4.7. Energy-band diagrams and current-voltage characteristic of a tunnel diode

(*a*) equilibrium state; (*b*) reverse bias; (*c*) region of rising tunneling current in the forward-biased state; (*d*) maximum tunneling current; (*e*) region of negative resistance; (*f*) rise in forward diffusion current

Such a form of the current-voltage characteristic, first described by the Japanese physicist L. Esaki in 1958, is associated with the flow of a tunneling current of majority carriers through a thin *p-n* junction.

In order that the tunneling current should flow through a *p-n* junction, the thickness of the junction must be sufficiently small (less than 100 Å) and on both sides of the junction there must be mutually equal energy levels between which the tunneling transition is made possible (see Sec. 2.6).

These conditions are fulfilled if the *p*- and *n*-regions are so strongly doped that the Fermi levels lie within the conduction band of the *n*-region and the valence band of the *p*-region (Fig. 4.7*a*). Such semiconductors are referred to as degenerate ones. The higher the degeneracy, the farther the Fermi levels, E_{Fp} and E_{Fn}, lie from the edges of the bands.

The tunneling current is proportional to the number of overlapping energy levels in the valence band and in the conduction band of the *p*- and *n*-type materials.

When the bias across the *p-n* junction is zero the energy levels do not overlap since the occupied levels in the *p*-region lie opposite the occupied levels in the *n*-region (Fig. 4.7*a*) and the current does not pass through the *p-n* junction.

If now the reverse bias is applied to this *p-n* junction (Fig. 4.7*b*), the Fermi level in the *p*-region will be above the Fermi level in the *n*-region by the value of external bias $V = (E_{Fp} + E_{Fn})/q$.

As a result, opposite the filled states in the *p*-region, which are below the Fermi level, there will appear empty states in the *n*-region, lying above the Fermi level, and the current will flow through the *p-n* junction (electrons will move from the valence band of the *p*-region and occupy the empty levels in the conduction band of the *n*-region). The higher the reverse bias, the larger the difference in the overlap of the levels and therefore the bigger the current (the value of current is illustrated by the width of the hatched portion in Fig. 4.7). Thus, the reverse tunneling current in the tunnel junction is very heavy, as distinct from the current in the common *p-n* junction. In the forward-biased junction the Fermi level becomes so shifted that opposite the vacant levels in the valence band of the *p*-region there develop filled levels in the conduction band of the *n*-region with the result that the tunneling current of electrons flows across the *p-n* junction from the *n*-region to the *p*-region (Fig. 4.7*c*).

This current will continue increasing to its highest value until the overlapping reaches a maximum (Fig. 4.7*d*).

The maximum overlap sets in when the forward voltage

$$V_p = \frac{E_{Fn} + E_{Fp}}{3q}$$

As the voltage is further increased the number of overlapping levels decreases (Fig. 4.7*e*) and at the voltage

$$V_v = \frac{E_{Fp} + E_{Fn}}{q}$$

this number goes to zero. The current flowing through the diode reduces now to a minimum. At voltages above V_v, the usual forward diffusion current passes through the *p-n* junction, which continues to grow with increasing voltage (Fig. 4.7*f*).

The current-voltage characteristic of a tunnel diode is shown in Fig. 4.7*g*. Here branch *I* is determined by an increasing number of overlapping levels in the *p*- and *n*-regions, while descending branch *II* results from a decrease in the band overlap with growing forward voltage. The second ascending branch *III* is due to a rise in the diffusion current.

The parameters of the tunnel diode are the peak voltage V_p, peak current I_p, valley voltage V_v, valley current I_v, and voltage V_{for} at which the diffusion current is equal to the peak current.

Voltages V_p and V_v depend on the values of E_{Fp} and E_{Fn}, i.e., on the degree of doping of the *p*- and *n*-regions of a semiconductor.

Valley current I_v, also called excess current, depends on the density of levels in a semiconductor band gap, which are responsible for the presence of the tunneling current at voltages higher than V_v. The larger the number of energy levels in the band gap, the higher the current I_v.

Voltage V_{for} depends on the width of a semiconductor band gap. Let us consider a *p-n* junction in which the *p*- and *n*-regions are so doped that their Fermi levels are in line with the edges of the bands (Fig. 4.8*a*). On applying a reverse bias to this junction, the electrons in the valence band of the *p*-region move on to the vacant levels in the conduction band of the *n*-region, causing a large tunneling current to flow through the *p-n* junction (Fig. 4.8*b*). In the forward direction, the levels do not overlap even under a small bias and the current does not flow through the *p-n* junction until after the appearance of a diffusion current of minority carriers in the forward-biased junction (Fig. 4.8*c*). The diodes with such a current-voltage characteristic are called backward as their forward current is smaller than the reverse one. The current-voltage characteristic of a backward diode is shown in Fig. 4.8*d*.

The speed of response of tunnel diodes is extremely high. The time of electron tunneling, $\tau = \varepsilon\varepsilon_0\rho$, comes to 10^{-13} s. This makes the negative resistance of the tunnel diode theoretically independent of the frequency of the voltage applied. The tunnel diodes actually operate with a slower speed of response. The limitation is due to the capacitance of the *p-n* junction and of the case.

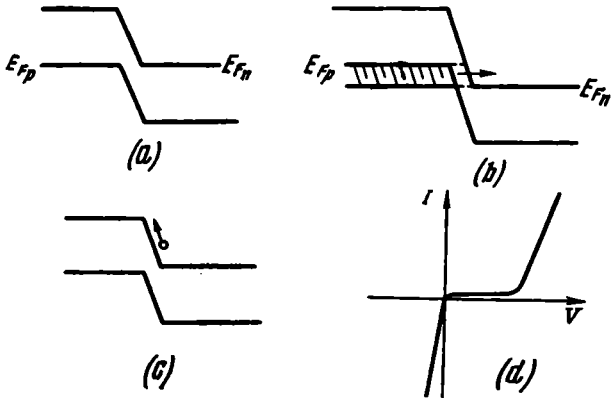


Fig. 4.8. Energy-band diagrams and current-voltage characteristic of a backward diode

(*a*) equilibrium state; (*b*) reverse bias; (*c*) forward bias; (*d*) current-voltage characteristic

The rate of variation in the voltage across the diode is determined by the ratio between the peak tunneling current and junction capacitance, I_p/C. This ratio is increased by making tunnel diodes from highly doped semiconductors (the dependence of the tunneling current on the impurity density is more pronounced than that of capacitance). The tunnel diode capacitance is also reduced by preparing the diodes in the form of abrupt fused *p-n* junctions 2 to 3 μm in diametre. To achieve the specified values of peak current I_p and reduce a junction area, diodes undergo electrolytic etching. Since the conductivity of a degenerate semiconductor is extremely high, the control of the value of I_p is done directly during etching which is terminated as soon as the required value of the peak current has been reached. Designs of some tunnel diodes are illustrated in Fig. 4.9.

Tunnel diodes can operate within a wide range of temperatures, which factor in combination with fast response offers considerable advantages when tunnel diodes are used as active elements in amplifiers, oscillators and logic circuits with a very small switching time.

A major property exhibited by the tunnel diode, allowing it to operate as an active element of the circuit, is a differential negative resistance which develops at voltage $V > V_p$. Its value is determined from the empirical formula

$$R_-[\text{ohm}] = \frac{140}{I_p} \quad \text{for Ge} \tag{4.20}$$

$$R_-[\text{ohm}] = \frac{200}{I_p} \quad \text{for GaAs} \tag{4.21}$$

The diode will function as an amplifier or oscillator when the condition $|R_-| > |R_l|$ is satisfied, otherwise the stability of the

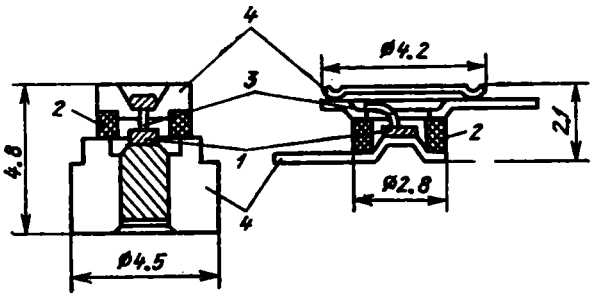


Fig. 4.9. Design of tunnel diodes
1—semiconductor crystal; 2—ceramic bushing; 3—contact wire; 4—leads

tunnel diode negative resistance will become dependent on the external resistance.

The cutoff frequency the diode operates at is found from the condition of equality to zero of the active component of the diode impedance and is described by the expressions

$$f=\frac{1}{2\pi R_{-}C}\sqrt{\frac{R_{-}}{r_r}-1} \quad \text{when} \quad R_{-}\geqslant 2r_r \tag{4.22a}$$

$$f=\frac{1}{4\pi r_r C} \quad \text{when} \quad R_{-}\leqslant 2r_r \tag{4.22b}$$

As is seen from (4.22), the frequency characteristics of tunnel diodes depend on the *p-n* junction capacitance and the base resistance.

As to their application, the tunnel diodes are subdivided into diode amplifiers, oscillators, and switches, differing in their parameters. Thus, the main parameter of the diode switch is the difference between the peak and the valley voltage ($V_v - V_p$) and also the ratio between the peak and the valley current (I_p/I_v), while that for the diode oscillator is the maximum power given up to the load.

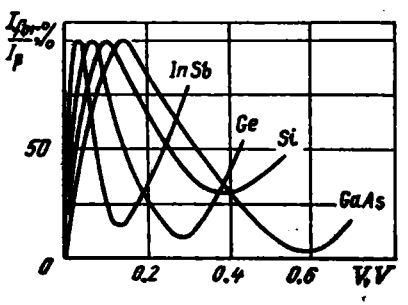


Fig. 4.10. Current-voltage characteristics of tunnel diodes made from different semiconductors

Since this power

$$P = 0.12\,(V_v - V_p)\,(I_p - I_v) \tag{4.23}$$

rises with increasing difference between V_v and V_p, it is advantageous to employ gallium arsenide diodes as oscillators and active elements of pulse circuits since they have the largest difference between the peak and valley voltage (see Fig. 4.10).

Diodes made of gallium antimonide (GaSb) have the lowest noise factor and are used in receiver-amplifier units operating at microwave frequencies. Yet the most popular are germanium tunnel diodes. Silicon tunnel diodes are impracticable as the ratio I_p/I_v of these devices never exceeds the figure of 3. The parameters of some tunnel diodes are tabulated in Tables 4.4 and 4.5.

Table 4.4

Parameters of Tunnel Diodes Designed to Operate in Pulse Circuits

Type	Material	Peak current I_p, mA	Peak-to-valley current ratio, I_p/I_v	Peak voltage, V_p, mV	Peak-valley voltage difference on current-voltage curve, $V_v - V_p$, mV
3И306	GaAs	2	8	170	850
ГИ304А	Ge	4.5	5	75	440
ГИ305А	Ge	9.1	5	81	450

Table 4.5

Parameters of Tunnel Diodes Designed for Operation in Amplifier and Oscillator Circuits

Type	Material	Peak current, I_p, mA	Peak-to-valley current ratio, I_p/I_v	Peak voltage, V_p, mV	Differential resistance, ohms
1И102А	Ge	1.25	5	70	75-110
3И101А	GaAs	1.0	5	160	24
АИ201А	GaAs	10	10	180	8

4.5. GUNN AND IMPATT DIODES

Gunn-effect diodes have recently found wide use in generating microwave-frequency oscillations. The Gunn effect shows up as high frequency oscillations that occur when the direct voltage applied

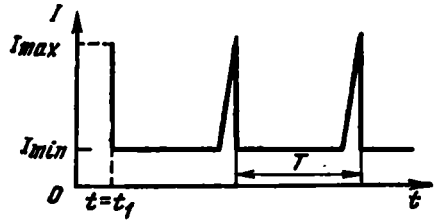


Fig. 4.11. Current pulse shape produced in a Gunn diode

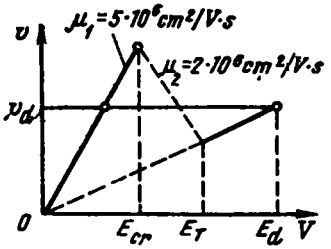


Fig. 4.12. Dependence of the electron velocity in gallium arsenide on the field strength

to a gallium arsenide monocrystal sets up an electric field of more than 10^3 V/cm.

The frequency of these oscillations is related to distance L between the contacts by

$$f = \frac{10^7}{L} \tag{4.24}$$

In testing specimens 100 μm long, the voltages at which oscillations occurred were in the order of several tens of volts and the oscillation frequency ran as high as 10^9 Hz, i.e., corresponded to the microwave range.

The emergence of spontaneous oscillations under the influence of high fields was also observed in monocrystals of some other semiconductors, such as indium phosphide InP, cadmium telluride CdTe, and gallium antimonide GaSb, but only gallium arsenide diodes have found extensive use so far.

The current pulse shape produced in the Gunn diode is shown in Fig. 4.11. Current oscillations occur because the mobility and speed of electrons diminish when they gain a critical energy. The electrons acquire this energy when the field strength exceeds the critical value.

Electron velocity v as a function of energy (field strength E) is represented in Fig. 4.12.

When $E < E_{cr}$, the current is caused to flow due to fast electrons with mobility $\mu_1 = 5 \times 10^3$ cm²/V·s, and when $E > E_{cr}$ the electrons move slower and their mobility $\mu_2 = 2 \times 10^3$ cm²/V·s.

As the current density $J = qnv$, the current-voltage characteristic displays a region of a differential negative resistance.

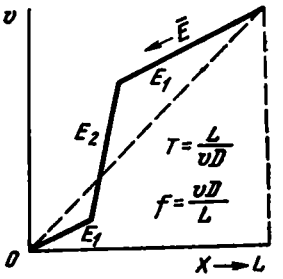


Fig. 4.13. Potential distribution along the length of a Gunn diode specimen

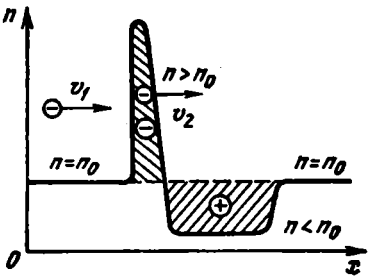


Fig. 4.14. Domain shape produced in a Gunn diode

The critical electric field strength usually develops near a negative electrode (cathode) of the diode. With the electron mobility reduced, the resistance rises in the region where the electrons have been slowed down, causing thereby a voltage redistribution in the specimen. The voltage drop and field strength increase in the high-resistance region and fall off in the remaining portion of the specimen.

The potential distribution along the length of the specimen is shown in Fig. 4.13.

In the high-field region the electron velocity is smaller than in those parts of the specimen where the electric field has not yet reached its critical value E_{cr}. Therefore, a negative space charge region is seen to adjoin the high-field region, as viewed from the cathode side, the charge being attributed to the fast electrons of high mobility overtaking this region. On the anode side, the high-field region borders the region of a positive space charge which has arisen in the area where fast electrons outrun the high-field region.

Thus, on both sides of the zone of low-speed electrons, there appears a double electric layer. This space charge layer containing heavy electrons of high density is called electric domain (Fig. 4.14).

As the strength of the electric field increases the field strength in the domain also increases until the carrier velocity inside the domain and that outside it become equal (point E_d in Fig. 4.12). Since this moment on, the field strength stops growing and the domain so formed moves to the anode at a constant speed v_d.

The domain size ranges from 0.1 to 0.03 of the specimen length and the domain field strength reaches 40 to 70 kV/cm.

The formation of the domain causes a decrease in the current to the value determined by the speed of slow electrons $J = qnv_d$.

The domain that has appeared near the cathode reaches the anode in time $t = L/v_d$ and is destructed there. The uniform field in the specimen is restituted and the current in it rises to a value of $J = qn\mu_1 E$. This gives rise to a new domain and thus to current oscillations in the circuit, the waveform of which is shown in Fig. 4.11

Let us consider the conditions under which the domain is formed For the domain instability to be preserved, the domain buildup time should be less than oscillation period T. This time is equal to dielectric relaxation time $\tau = \varepsilon\varepsilon_0\rho$.

Hence, the domain instability is ensured when $T > \varepsilon\varepsilon_0\rho$, or

$$T > \frac{\varepsilon\varepsilon_0}{qn_0\mu}$$

Since

$$T = \frac{L}{v_d}, \quad \frac{L}{v_d} > \frac{\varepsilon\varepsilon_0}{qn\mu}$$

then

$$nL > \frac{\varepsilon\varepsilon_0 v_d}{q\mu} \tag{4.25}$$

When $\varepsilon = 12.5$, $v_d = 10^7$ cm/s and $\mu = 5000$ cm²/V·s,

$$nL > 1.7 \times 10^{13} \text{ cm}^{-2} \tag{4.26}$$

Gunn diodes are usually made from monocrystal bars of gallium arsenide with a carrier density of 10^{13} to 10^{15} cm^{-3} or epitaxial films of GaAs grown on high-resistivity gallium-arsenide crystals. The use of epitaxial films makes possible the preparation of small-size diodes and even microwave integrated circuits by planar methods

The working frequencies of Gunn oscillators, as is seen from formula (4.24), go up to 10^{10} Hz. At a crystal size less than 10 μm, inhomogeneity in the crystal structure begins to show up and stable oscillations become difficult to produce.

These difficulties are easily circumvented if the Gunn diodes are made to operate in the so-called limited space-charge accumulation mode (LSA mode). When in this mode of operation, the Gunn diode connected into an oscillatory circuit is fed with a steady bias voltage V_b and an alternating-current voltage of such values that during a portion of the period the field strength across the diode becomes less than the critical value (hatched portions in Fig. 4.15).

So long as the diode field strength $E = V/L$ is above the critical value, a domain is being formed near the cathode. However, if the

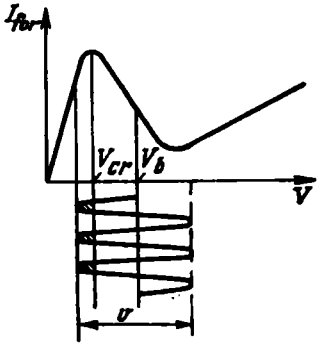


Fig. 4.15. Current-voltage characteristic of a Gunn diode operating in the LSA mode

frequency of the alternating voltage is sufficiently high, the domain which has not yet formed begins to die out when the field strength becomes less than E_{cr} (when $V < V_{cr}$).

The current through the diode diminishes during the buildup of the domain and rises when the latter gets extinguished, in other words, the diode exhibits a negative resistance, and for this reason the circuit with the diode connected into it is able to produce electric oscillations. The oscillation period maintained under the LSA mode of operation is determined not by the domain transit time $t = L/v_d$, as is the case with the Gunn mode, but by the resonator frequency.

In theory, the oscillators of this type can operate at a frequency of 30 to 300 GHz, releasing pulse energies in the order of hundreds of kilowatts at an efficiency of 20%. Known are Gunn oscillators which at a frequency of 1 GHz have a pulse power of 1000 W; at GHz, 360 W; and at 10 GHz, 33 W.

The Gunn effect has already found practical application in miniature Doppler radars and navigation systems and also in phased-array elements of radars.

Impact avalanche-and-transit time diodes (IMPATT diodes) are the devices which also have a negative resistance but, unlike the Gunn oscillators and tunnel diodes, they exhibit negative resistance only when fed with a high-frequency alternating voltage. The operation of IMPATT diodes is based on the mechanism of breakdown in a *p-i-n* junction.

The electron-hole pairs generated during the breakdown process move to space charge regions in a high electric field. When the field strength exceeds 50 kV/cm, the electron speed ceases to rise and becomes constant with increasing field strength. This saturation of the electron speed causes the phase shift of the current with respect to the high-frequency alternating voltage applied to the diode.

A phase difference of 180° can be achieved by selecting an appropriate value of the applied voltage and the width of the space charge

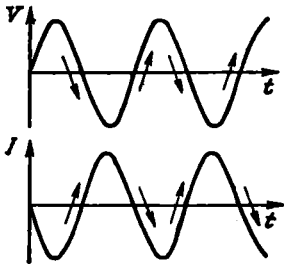


Fig. 4.16. Development of a negative resistance due to the phase shift between the current and voltage in an IMPATT diode (180° out-of-phase on the drawing)

region. In this case the reduction in the applied voltage will correspond to the rise in the current (Fig. 4.16), i.e., at a high frequency the diode will have a negative resistance. The operating frequency of IMPATT diodes depends on the width of the space charge region. The narrower the region, the higher the operating frequency. This is why IMPATT diodes are made not only with p^+-i-n junctions, but also with p-n junctions.

IMPATT diodes are manufactured from germanium, silicon and gallium arsenide. They operate at microwave frequency ranges up to 50 GHz and are rated up to 350 mW.

Disadvantages of these diodes are a high noise level and a low efficiency (1 to 10%), the latter shortcoming being attributed to a rather narrow allowable range of voltages across the diode, as, on the one hand, the voltage should not exceed too much the critical breakdown voltage and, on the other hand, it should not be too low, otherwise the stability condition of the drift velocity becomes disturbed.

Review Questions

1. Why does the reverse-resistance recovery time depend on the forward current?
2. Why does the amplitude of the current through a diode vary when the pulse duration is below the reverse-resistance recovery time?
3. What physical phenomena determine the values of t_1 and t_2 in switching diodes?
4. What is the after-injection emf?
5. What criteria determine the regulation voltage across a Zener diode and how can one control it?
6. What physical processes govern the TCV of a tunnel and an avalanche breakdown?
7. What determines the speed of response of varicaps?
8. What is the difference between MIS capacitors and varicaps with p-n junctions?
9. What is the excess current in tunnel diodes and what determines its value?

Chapter Five

TRANSISTOR: PRINCIPLES AND CHARACTERISTICS

5.1. TRANSISTOR AS AN AMPLIFIER

The transistor is a semiconductor device made up of two *p-n* junctions in a single semiconductor crystal and intended to amplify, generate and switch over electric currents.

The alternation of regions of different conductivity type can be made to follow either a *p-n-p* or an *n-p-n* pattern (Fig. 5.1). The principle of operation of *p-n-p* and *n-p-n* transistors is the same, the difference being in the sign of external sources of voltage applied and in the direction of the current flow through electrodes. Therefore, the conclusions to be drawn in this Chapter for *p-n-p* transistors are quite applicable for *n-p-n* transistors.

The middle region of a transistor (*n*-type region for a *p-n-p* transistor) is the base and the extreme regions lying on both sides of the base are correspondingly called the emitter and the collector region, while the *p-n* junctions they form with the base region are referred to as the emitter and collector junctions, respectively (Fig. 5.1). When the transistor is used as an amplifier, the external sources of voltage are so connected that the emitter junction is biased in the forward direction and the collector junction, in the reverse direction (Fig. 5.2).

When in the forward-biased state, the emitter junction injects minority carriers (holes) into the base region, wherefrom the holes diffuse in the direction from the emitter region with a lower density to the collector region with a higher density.

Since the base region width W is specially selected such as to be smaller than the diffusion length L_p of minority carriers, the recombination process in the base is insignificant and most of the injected holes reach the collector. The current in the output circuit will be thus almost equal to the emitter current. A voltage drop develops ac-

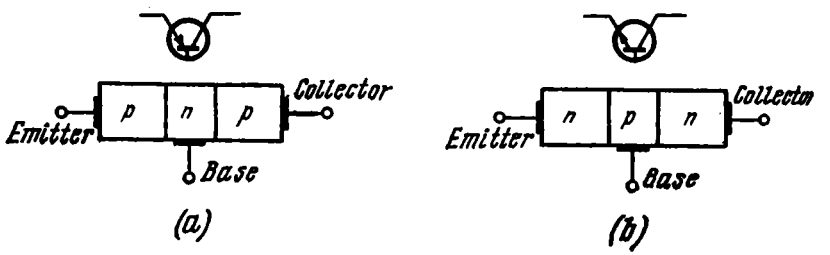


Fig. 5.1. Structure of a *p-n-p* (*a*), and *n-p-n* (*b*) transistor

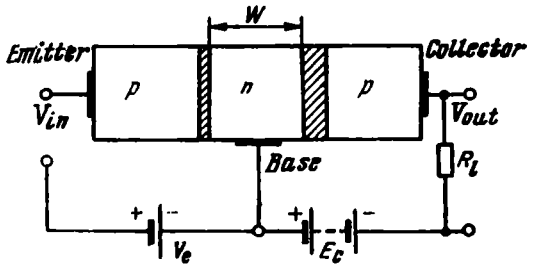


Fig. 5.2. A *p-n-p* transistor connection diagram

ross a load resistance R_l series-connected with the collector junction, the voltage drop being proportional to the collector current through the resistance.

The number of holes that reach the collector is determined by the number of injected holes which are dependent on the magnitude of the voltage across the emitter, i.e., the collector current is controlled by the emitter current. The power gain in the transistor is enabled due to the fact that the emitter and collector currents are almost equal, while the voltage across the load resistance is appreciably higher than the voltage across the low-resistance emitter junction biased in the forward direction.

It is of importance for the power gain mode being considered here that the voltage drop across the load resistance should not exceed the voltage applied to the collector *p-n* junction, in other words, that the following condition be observed:

$$I_c R_l < E_c \tag{5.1}$$

If this condition is not fulfilled the collector will become forward biased and instead of drawing holes by its field, which have been injected by the emitter into the base, it will start to inject holes.

We have considered the amplification of signals fed to the emitter in the forward direction. For the transistor to amplify signals of opposite sign without distortion, a constant forward bias is commonly impressed on the emitter-base junction. In this case the transistor responds to a back bias signal by reducing the collector current and changing the output signal across the load resistance. If the transistor input is supplied with a signal of alternating polarity, an amplified signal of the same shape may be derived at the output of the transistor, i.e., in the collector circuit. The current bias is applied on the emitter in the similar way as the voltage bias on the control grid of an electronic valve.

5.2. CALCULATION OF CURRENT GAIN

One of the main characteristics of a transistor is a current gain factor h_{21b} or α (alpha current gain), defined as

$$h_{21b} = \left(\frac{\partial I_c}{\partial I_e}\right)_{V_c=\text{const}} \tag{5.2}$$

It characterizes a change in the output current I_c with a change in the input current I_e at a constant collector voltage V_c.

Let us determine the values of currents flowing through the transistor electrodes in the circuit of Fig. 5.2. We assume that $p \gg n$ in the emitter and collector regions and that the density distribution of holes p and electrons n in the transistor corresponds to that shown in Fig. 5.3. It may also be assumed that the density of hole current[1] I_p throughout the base region is uniform and equal to the density of emitter hole current I_{pe}, because the base region width W is usually made smaller than the diffusion length of holes L_p in order to reduce recombination losses. Practically, all the applied voltage drops off across the collector; therefore the field of the base may be ignored in the calculations and the current in the base region may be considered to be a purely diffusion current.

We write down I_{pe} in the form

$$I_{pe} = -qD_p \frac{dp}{dx} \tag{5.3}$$

Dividing the variables and performing the integration, we obtain the expression for the hole density in the base region (x varies from 0 to W)

$$p(x) = -\frac{I_{pe}x}{qD_p} + p(0) \tag{5.4}$$

Knowing from (2.18) that when $x = 0$

$$p(0) = p_n e^{\frac{qV_e}{kT}} \tag{5.5}$$

and when $x = W$

$$p(W) = p_n e^{-\frac{qV_c}{kT}} \tag{5.6}$$

we find from (5.4)

$$I_{pe} = \frac{qD_p p_n}{W}\left(e^{\frac{qV_e}{kT}} - e^{-\frac{qV_c}{kT}}\right) \tag{5.7}$$

[1] In our further discussion by currents identified with symbol I will be understood the currents flowing through a unit cross-sectional area, i.e., the current density.

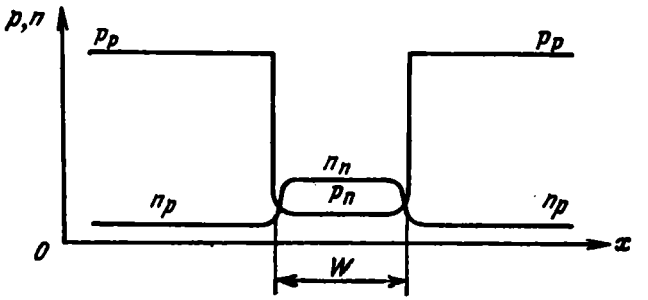


Fig. 5.3. Charge-carrier density distribution in a *p-n-p* transistor

When $V_c \gg kT/q$, which is the case for the transistor in operation, the second term in (5.7) may be disregarded. Hence,

$$I_{pe} = \frac{qD_p p_n}{W} e^{\frac{qV_e}{kT}} \tag{5.8}$$

Substituting (5.5) and (5.8) into (5.4) yields

$$p(x) = p_n e^{\frac{qV_e}{kT}} \left(1 - \frac{x}{W}\right) \tag{5.9}$$

A certain amount of holes out of the total emitter current of holes I_{pe} recombine in the base region and consequently the hole current near the collector

$$I_{pc} = I_{pe} - I_{rec} \tag{5.10}$$

where I_{rec} is the recombination component of the emitter hole current.

Let us estimate the value of I_{rec}.

The time variation in the hole density due to hole recombination in the bulk of the base obeys the following equation:

$$\left(\frac{dp}{dt}\right)_{rec} = -\frac{p - p_n}{\tau_p} \tag{5.11}$$

where τ_p is the lifetime of holes in the base region. Then,

$$I_{rec} = qS_e \int_0^W \frac{p - p_n}{\tau_p} dx \tag{5.12}$$

The substitution of (5.9) into (5.12) and subsequent integration yield

$$I = \frac{qS_e}{\tau_p} \int_0^W \left[p_n e^{\frac{qV_e}{kT}} \left(1 - \frac{x}{W}\right) - p_n\right] dx = \frac{qS_e p_n W}{2\tau_p} \left(e^{\frac{qV_e}{kT}} - 2\right) \tag{5.13}$$

Bearing in mind that at room temperature $qV_e/kT \gg 1$, for the values of V_e exceeding one tenth of a volt we obtain

$$I_{rec} = \frac{qS_e p_n W}{2\tau_p} e^{\frac{qV_e}{kT}} \tag{5.14}$$

The transport factor β^* that characterizes the quantity of holes passing through the base region, and is expressed as a ratio between the collector and emitter hole currents

$$\beta^* = \frac{I_{pc}}{I_{pe}} \tag{5.15}$$

will be expressed in the form

$$\beta^* = \frac{I_{pc}}{I_{pe}} = 1 - \frac{I_{rec}}{I_{pe}} = 1 - \frac{W^2}{2D_p\tau_p} \tag{5.16}$$

or, considering that $\sqrt{D_p\tau_p} = L_p$, in the form

$$\beta^* = 1 - \frac{1}{2}\left(\frac{W}{L_p}\right)^2 \tag{5.17}$$

The electron component of the emitter current near the base region boundary, as follows from (2.18), is equal to

$$I_{ne} = \frac{qS_e D_n n_p}{L_n}\left(e^{\frac{qV_e}{kT}} - 1\right) \approx \frac{qS_e D_n n_p}{L_n} e^{\frac{qV_e}{kT}} \tag{5.18}$$

because $qV_e \gg kT$. The net emitter current I_e equals the sum of hole and electron components given in (5.8) and (5.18), respectively

$$I_e = I_{pe} + I_{ne} = e^{\frac{qV_e}{kT}}\left(\frac{qD_p p_n}{W} + \frac{qD_n n_p}{L_p}\right) S_e \tag{5.19}$$

Let us introduce the notion of the emitter injection efficiency γ^* which is the ratio of the hole component I_{pe} of the emitter current to the net emitter current and which is expressed as

$$\gamma^* = \frac{I_{pe}}{I_e} = \frac{1}{1 + (I_{ne}/I_{pe})} = \frac{1}{1 + (D_n n_p W / D_p p_n L_n)} \tag{5.20}$$

According to (2.1), $n_p/p_n = n_n/p_p$, and hence expression (5.20) may be written in the form

$$\gamma^* = \frac{1}{1 + (W\sigma_b/L_n\sigma_e)} \approx 1 - \frac{\sigma_b}{\sigma_e}\frac{W}{L_n} \tag{5.21}$$

where σ_b and σ_e are the conductivities of the base and emitter regions, respectively, and $\sigma_b W/\sigma_e L_p \ll 1$ since $W/L_p < 1$ and $p_p/n_n \ll 1$.

It is evident from the equation that the higher the emitter conductivity compared to that of the base and the thinner the base as compared with the diffusion length of electrons, the higher the emitter efficiency.

In deriving expression (5.16) for the transport efficiency, account was taken of only the bulk recombination. Recombination on the surface can materially affect the parameters of a transistor. It is characterized by the surface recombination velocity s. The value of s can vary depending on the technique of surface treatment in the process of the manufacture of a transistor. Whatever the attempts to minimize the value of s, a certain amount of holes do not reach the collector junction due to surface recombination and thus it must be taken into account.

The velocity of recombination on a semiconductor surface is determined as a ratio of the density of carriers being recombined per unit surface area per second to the excess carrier density at the surface. To obtain the value of the current with due regard for surface recombination (variations in the charge per unit time), the value of s should be multiplied by the electron charge q, effective surface area S_s over which the recombination takes place, and also by the carrier density.

As the real emitter junction area is usually smaller than the collector junction area and, besides, the current density across the emitter exceeds the collector current density, the surface recombination may be thought to take place mainly near the emitter junction where $p_e = p_n e^{qV_e/kT}$. The surface recombination current may therefore be written in the form

$$I_{rec\,s} = qsS_s p_n e^{\frac{qV_e}{kT}} \tag{5.22}$$

Dividing (5.22) by (5.8) yields

$$\frac{I_{rec\,s}}{I_{pe}} = \frac{qsS_s p_n e^{\frac{qV_e}{kT}}}{\frac{qD_p p_n}{W} e^{\frac{qV_e}{kT}} S_e} = \frac{sS_s W}{D_p S_e} \tag{5.23}$$

Where S_e is the emitter junction area.

The transport factor β^*, with regard for the bulk and surface recombination, will be consequently expressed in the form

$$\beta^* = 1 - \frac{1}{2}\left(\frac{W}{L_p}\right)^2 - \frac{sS_s W}{D_p S_e} \tag{5.24}$$

From (5.24) it follows that in order to increase β^*, i.e., the portion of the minority carriers reaching the collector, it is required to reduce both the ratio of the base region width to the diffusion length of the minority carriers and the coefficient of surface recombination.

An electron component I_{nc} of the collector current is produced by electrons passing from the collector into base. At the base-collec-

tor boundary the density of electrons is

$$n_{bc} = n_p e^{-\frac{qV_c}{kT}} \tag{5.25}$$

and, by analogy with (5.18), the current they carry is given by

$$I_{nc} = \frac{qS_c D_n n_p}{L_n}\left(e^{-\frac{qV_c}{kT}} - 1\right) \approx -\frac{qS_c D_n n_p}{L_n} \tag{5.26}$$

The comparison of (5.26) with (5.18) shows that $I_{nc} \ll I_{ne}$ because $V_e \gg kT/q$ and for this reason the total collector current may be presumed to be essentially equal to hole current I_{pc}. It is seen from (5.16) and (2.20) that

$$I_c = I_{pc} = \beta^* I_{pe} = \beta^* \gamma^* I_e = \alpha I_e \tag{5.27}$$

Let us note that by Kirchhoff's law the base current

$$I_b = I_e - I_c = I_e(1 - \alpha) \tag{5.28}$$

It is evident from (5.27) that for a *p-n-p* transistor the current gain $\alpha = \beta^* \gamma^*$ or

$$\alpha = \left[1 - \frac{1}{2}\left(\frac{W}{L_p}\right)^2 - \frac{sS_s W}{D_p S_e}\right]\left(1 - \frac{\sigma_b W}{\sigma_e L_n}\right) = 1 - \frac{1}{2}\left(\frac{W}{L_p}\right)^2 - \frac{sS_s W}{D_p S_e} - \frac{\sigma_b W}{\sigma_e L_n} + \frac{1}{2}\frac{\sigma_b W^3}{\sigma_e L_p^2 L_n} + \frac{sS_s W^2}{S_e D_p \sigma_e L_n} \tag{5.29}$$

Since $W \ll L_n$, the last two terms may be omitted and then we finally have

$$\alpha \approx 1 - \frac{\sigma_b W}{\sigma_e L_n} - \frac{1}{2}\left(\frac{W}{L_p}\right)^2 - \frac{sS_s W}{D_p S_s} \tag{5.30}$$

Two important conclusions can be drawn from this expression:

1. The alpha current gain is always less than unity.
2. To obtain the maximum value of α for a transistor, one should (a) reduce the base width, (b) increase the lifetimes of carriers in the base, (c) reduce the surface recombination velocity, and (d) increase the ratio between the conductivities of the emitter and the base region.

5.3. INTRINSIC RESISTANCES OF THE TRANSISTOR

The most convenient configuration of the equivalent network for considering the operation of a transistor is the T equivalent circuit shown in Fig. 5.4. The circuit includes three resistances, namely, emitter junction resistance r_e, collector junction resistance r_c, and base region resistance r_b, which for bravity will be referred to as the emitter, collector, and base resistance, respectively. Equivalent

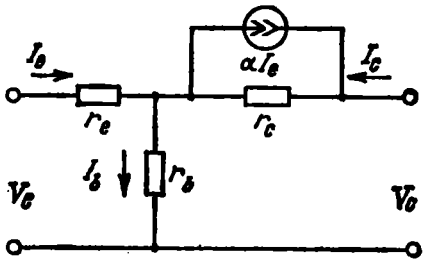


Fig. 5.4. T equivalent circuit for a low-frequency transistor

current generator αI_e connected in parallel with resistance r_c and setting current independent of the load displays the properties of a transistor as an active element of the circuit.

The emitter resistance is given by

$$r_e = \left(\frac{\partial V_e}{\partial I_e}\right)_{I_c=\text{const}} \tag{5.31}$$

Differentiating expression (5.19) for I_e with respect to V_e

$$\frac{\partial I_e}{\partial V_e} = \frac{q}{kT}\left(\frac{qD_p p_n}{W} + \frac{qD_n n_p}{L_n}\right) e^{\frac{qV_e}{kT}} = \frac{q}{kT} I_e \tag{5.32}$$

we find r_e

$$r_e = \frac{kT}{qI_e} \tag{5.33}$$

The collector resistance is accordingly equal to

$$r_c = \left(\frac{\partial V_c}{\partial I_c}\right)_{I_e=\text{const}} \tag{5.34}$$

Since

$$q_c = \frac{1}{r_c} = \left(\frac{\partial I_c}{\partial V_c}\right)_{I_e=\text{const}}$$

then, setting $I_c = \alpha I_e$ and considering that α depends on the collector voltage, it is easy to obtain

$$q_c = I_e S_c \frac{\partial \alpha}{\partial V_c} = I_e S_c \frac{\partial \alpha}{\partial W}\frac{\partial W}{\partial V_c} \tag{5.35}$$

By using (5.30), the ratio $\partial\alpha/\partial W$ may be expressed in the form

$$\frac{\partial \alpha}{\partial W} = -\frac{\sigma_b}{\sigma_e L_n} - \frac{1}{2}\frac{2W}{L_p^2} - \frac{sS_e}{D_p S} =$$
$$= -\frac{1}{W}\left\{\frac{\sigma_b W}{\sigma_e L_n} + 2\left[\frac{1}{2}\left(\frac{W}{L_p}\right)^2\right] + \frac{sS_e W}{D_p S}\right\} \tag{5.36}$$

The substitution of the transport factor β^* from (5.24) and emitter efficiency γ^* from (5.21) brings formula (5.36) to the final form

$$\frac{\partial \alpha}{\partial W} = -\frac{1}{W}\left[(1-\gamma^*) + 2(1-\beta^*)\right] \tag{5.37}$$

The base width changes due to a change in the space charge near the collector junction. Therefore,

$$\frac{\partial W}{\partial V_c} = \frac{\partial W}{\partial \delta_c} \frac{\partial \delta_c}{\partial V_c} \tag{5.38}$$

where δ_c is the thickness of a depletion layer in the collector junction. Since for the case being considered $\delta_c \gg \delta_b$, the expansion of the space charge layer in the collector region may be neglected on the assumption that the collector depletion layer is disposed almost completely in the base region. This means that a decrease in the base width is equal to an increase in the space-charge layer width, i.e.,

$$\frac{\partial W}{\partial \delta_c} = -1 \tag{5.39}$$

As shown in (2.16)

$$\delta_c = \sqrt{\frac{2\varepsilon\varepsilon_0 (V_c + \Delta\varphi_c)}{q n_n}}$$

therefore,

$$\frac{\partial \delta_c}{\partial V_c} = \frac{\delta_c}{2V_c} \tag{5.40}$$

We thus find from (5.37), (5.39) and (5.40) that

$$\frac{\partial I_c}{\partial V_c} = I_e \frac{\delta_c}{2V_c} \frac{1}{W} [2(1-\beta^*) + (1-\gamma)] \tag{5.41}$$

Accordingly,

$$r_c = \frac{2WV_c}{\delta_c [2(1-\beta^*) + (1-\gamma)] I_e} \tag{5.42}$$

The analysis of (5.42) yields the following:

1. The collector resistance becomes greater with increasing base region width W. At the same time one must bear in mind that an increase in W leads to a reduction of β^* due to a rise in the recombination losses.
2. The collector resistance increases with a decrease in the collector junction width δ_c, i.e., with a rise in the density of majority carriers in the base region (electrons in our case). But an excessive increase of n_n causes the electron component of the emitter current to rise and the emitter efficiency γ to fall.
3. Since δ_c is proportional to $\sqrt{V_c}$, r_c is also proportional to $\sqrt{V_c}$.
4. To obtain a high output resistance of the transistor, it is more preferable that the transistor should operate in the range of small emitter currents.

The base resistance r_b may be regarded as a sum of ohmic resistance r_{b1} and diffusion resistance r_{b2} of the base region

$$r_b = r_{b1} + r_{b2} \tag{5.43}$$

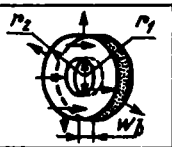
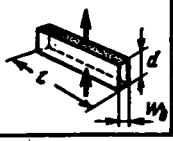
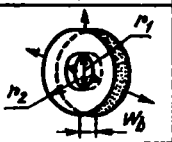
element shape	resistance	element shape	resistance
	$r_{b_1} = \frac{\rho}{3W_b}\frac{h}{l}$		$r_{b_1} = \frac{\rho}{8\pi W_b}$
	$r_{b_1} = \frac{\rho}{12W_b}\frac{h}{l}$		$r_{b_1} = \frac{\rho}{4\pi W_b}\ln\frac{r_2}{r_1}$
	$r_{b_1} = \frac{\rho}{W_b}\frac{d}{l}$		$r_{b_1} = \frac{\rho}{2\pi W_b}\ln\frac{r_2}{r_1}$

Fig. 5.5. Summary table of approximate formulas for calculating the bulk resistance of base elements of different configurations (arrows indicate the direction of current flow)

The ohmic resistance of the base depends on the resistivity of the base material and configuration of the base. Fig. 5.5 gives formulas for the calculation of r_{b1} for variously shaped junction transistors met with in practice. More complicated base configurations may be calculated as a sum of several portions illustrated in Fig. 5.5. The resistance r_{b2} characterizes the effect of collector voltage V_c on base width W due to a change in the space-charge layer width δ_c. Let us illustrate this by an example. Let W' correspond to a collector voltage V_c (Fig. 5.6). Assuming $\alpha = 1$, we have a steady-state gradient of the hole density. As the voltage across the collector increases to V_c'', the base width becomes equal to W'' according to (2.16), but

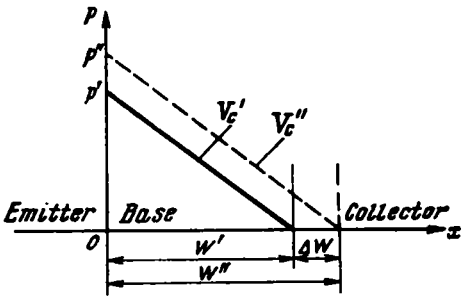


Fig. 5.6. Variations in the density distribution of holes in the base region with changing voltage across the collector

the hole density gradient does not change since by formula (5.3) dp/dx = const at a constant emitter current. As is seen from the drawing, the hole density p near the emitter must diminish, which is possible at the invariable emitter current only due to a decrease in the voltage across the emitter junction, this being equivalent to the appearance of an additional base resistance r_{b2} not associated with the resistivity of the base material. The expression for r_{b2} may be written in the form

$$r_{b2}' = \left(\frac{\partial V_e}{\partial I_c}\right)_{I_b=\text{const}} \tag{5.44}$$

Let us introduce this expression in the following way:

$$r_{b2} = \frac{\partial V_e}{\partial V_c}\frac{\partial V_c}{\partial I_c} = \frac{\partial V_e/\partial V_c}{\partial I_c/\partial V_c} \tag{5.45}$$

It is more convenient to write down the numerator as a derivative of the complex function

$$\frac{\partial V_e}{\partial V_c} = \frac{\partial V_e}{\partial p_{be}}\frac{\partial p_{be}}{\partial W}\frac{\partial W}{\partial V_c} \tag{5.46}$$

where p_{be} is the hole density at the base-emitter boundary. Let us find each of the derivatives. Since

$$p_{be} = p_n e^{\frac{qV_e}{kT}}$$

then

$$\frac{\partial V_e}{\partial p_{be}} = \frac{kT}{qp_{be}} \tag{5.47}$$

As has been shown above, the hole density gradient is constant and hence,

$$\frac{\partial p_{be}}{\partial W} = \text{const} = -\frac{p_{be}}{W} \tag{5.48}$$

Substituting (5.47) and (5.48) into (5.46) yields

$$\frac{\partial V_e}{\partial V_c} = \frac{kT}{qW}\frac{\partial W}{\partial V_c} \tag{5.49}$$

Bearing in mind that $I_c = \alpha I_e$, one may write

$$\left(\frac{\partial I_c}{\partial V_c}\right)_{I_e=\text{const}} = I_e\frac{\partial \alpha}{\partial W}\frac{\partial W}{\partial V_c} \tag{5.50}$$

Dividing (5.49) by (5.50) and substituting $\partial\alpha/\partial W$ from (5.37), we obtain

$$r_{b2} = \frac{kT}{qI_e\,[2(1-\beta^*)+(1-\gamma^*)]} \tag{5.51}$$

Considering (5.33), the expression for r_{b2} will be finally written as

$$r_{b2} = \frac{r_e}{2(1-\beta^*)+1-\gamma^*} \tag{5.52}$$

Thus, at not extremely high current densities one component r_{b1} of the base resistance is determined only by the geometry of the transistor and the material resistivity. The other component r_{b2} varies directly with the emitter resistance, is related to the components of the current gain and is heavily dependent on the emitter current and temperature.

5.4. TRANSISTOR CURRENTS AND VOLTAGES

The collector current may be regarded as a sum of two current components, the first being attributed to the carrier injection across the emitter junction and the second, to the reverse bias on the collector:

$$I_c = \alpha I_e + I_{c0} \tag{5.53}$$

where I_{c0} is the current in the reverse-biased collector junction with the emitter circuit open ($I_e = 0$).

According to (2.32) one may write

$$I_{c0} = Sq\left(\frac{D_p p_n}{L_p} + \frac{D_n n_p}{L_n}\right)\left(e^{-\frac{qV_c}{kT}} - 1\right) = I_0\left(e^{-\frac{qV_c}{kT}} - 1\right) \tag{5.54}$$

Generally $V_c \gg kT/q$ and it may be assumed therefore that I_{c0} does not depend on voltage within an appreciably wide range up to the breakdown point and is equal to saturation current I_0. In our case, when $n_p \ll p_n$ the collector reverse current will be mainly determined by the holes being transferred from the n-region into the p-region

$$I_{c0} = qD_p \frac{p_n}{L_p} S = q\frac{n_i^2}{n_n}\frac{D_p}{L_p} S \tag{5.55}$$

Thus, the quantity I_{c0} is significantly dependent on the properties of the semiconductor material employed (n_n, L_p) and temperature. To provide for a dependable operation of the transistor, it is essential that the value of I_{c0} be reduced to a minimum.

In real transistors, surface conditions exert a substantial influence on I_{c0}. Moisture and contaminants appreciably increase I_{c0} and thus cause shunting of a p-n junction and give rise to surface leakage.

The collector junction breakdown results, as a rule, from collision ionization and resembles an identical phenomenon occurring in diodes, which has been described in detail in Chapter Two.

The avalanche multiplication being taken into account, the collector current for the case of an open emitter circuit ($I_e = 0$) may

be written as

$$I_c = I_{c0}M = \frac{I_{c0}}{1-(V_c/V_a)^n} \tag{5.56}$$

where M is the avalanche multiplication factor.

Under the breakdown condition $I_c \to \infty$, which is possible when $V_c = V_a$.

When the base circuit is open, i.e., when $I_b = 0$ and consequently $I_c = I_e$, formula (5.53) can be written as $I_c(1 - \alpha) = I_{c0}$. With the avalanche multiplication taking place, the current gain factor α_M increases M times as compared with the current gain α under normal conditions. Current I_{c0} also rises M times. Hence,

$$I_c = \frac{I_c M}{1-\alpha M} \tag{5.57}$$

Breakdown occurs when $I_c \to \infty$, i.e., when $\alpha M = 1$. Since $M = 1/\left[1 - \left(\frac{V_c}{V_a}\right)^n\right]$, then

$$V_c = V_a \sqrt[n]{1-\alpha} \tag{5.58}$$

It follows from the above that in the case of an open base circuit the collector breakdown voltage is by a factor of $\sqrt[n]{1-\alpha}$ smaller than in the case of an open emitter circuit.

Quantity α depends on the emitter current, for at a high current density recombination losses grow. Similarly, quantity V_c at which breakdown occurs between the emitter and collector depends in the general case on the emitter current and is designated therefore as V_a. Generally speaking, at $I_b \neq 0$ the emitter-collector breakdown voltage depends on the emitter-base circuit resistance and tends to diminish with an increase of the latter. As has been stated above, V_a is minimum when $R_b \to \infty$ ($I_b = 0$).

The reasonings pertaining to the breakdown of the reverse-biased collector junction can be equally applied to the emitter junction when the latter is supplied with the reverse bias voltage.

Sometimes a punch-through of p-n junctions is observed in the operating transistor. The voltage at which this effect takes place can appreciably limit the range of working voltages even in the region remote from the breakdown point.

As is known, with increasing reverse voltage the width of the collector junction grows, according to (2.16), in proportion to $V_c^{1/2}$. By varying V_c, the value of δ_c may be brought to the value of the base region width W. In this case the collector junction reaches the boundary of the emitter junction and the current in the emitter-collector circuit sharply rises. The punch-through voltage is calculated by the formula

$$V_p = \frac{W^2}{2\varepsilon\varepsilon_0\mu_b} \tag{5.59}$$

For a complex-shaped base region, W should be understood to be a minimum value of the base width. The punch-through effect can be observed only when the value of V_p is below V_a.

5.5. TRANSISTOR CAPACITANCE

Barrier capacitances of reverse-biased emitter and collector junctions are similar to capacitances in individual p-n junctions. Therefore, the calculations of capacitances for abrupt p-n junctions are made by using formulas analogous to formula (2.53):

$$C_{cb} = S_c \sqrt{\frac{\varepsilon\varepsilon_0 qN}{2(\Delta\varphi_c - V_c)}} \tag{5.60}$$

$$C_{eb} = S_e \sqrt{\frac{\varepsilon\varepsilon_0 qN}{2(\Delta\varphi_c - V_e)}} \tag{5.61}$$

However, diffusion capacitances of p-n junctions also have a pronounced effect on the operation of a transistor. The emitter-junction diffusion capacitance is conditioned by a variation in the charge of nonequilibrium carriers with a change in the voltage applied to the emitter junction. With an increase of the forward bias on the emitter, holes are injected into the base and as soon as the bias drops off, the holes are drawn from it. A change in the space charge of holes is compensated for by the inflow of electrons through the base contact. Thus, the variations in the voltage result in the variations of the base charge, which is identical to the presence of a capacitance. The emitter diffusion capacitance is

$$C_{e\,dif} = \frac{\partial Q}{\partial V_e} = q\frac{\partial P}{\partial V_e} \tag{5.62}$$

and

$$P = S\int_0^W [p(x) - p_n]\,dx \tag{5.63}$$

where P is the total number of excess holes in the base region.

Using expression (5.14), we transform formula (5.63) to the following form

$$P = \frac{I_{rec}\tau_p S}{q} = \frac{p_n WS}{2} e^{\frac{qV_e}{kT}} \tag{5.64}$$

Hence,

$$C_{e\,dif} = q\frac{\partial P}{\partial V_e} = \frac{q^2 WS}{2kT} p_n e^{\frac{qV_e}{kT}} \tag{5.65}$$

Substituting

$$p_n e^{\frac{qV_e}{kT}} = \frac{WI_e}{qD_p}$$

into (5.65) from (5.8) and using (5.33), we obtain a relation between the emitter diffusion capacitance and the emitter current or resistance:

$$C_{e\,dif} = \frac{qW^2}{2kTD_p} I_e = \frac{W^2}{2D_p} \frac{1}{r_e} \tag{5.66}$$

The total emitter capacitance C_e is equal to

$$C_e = C_{eb} + C_{e\,dif} \tag{5.67}$$

where $C_{e\,dif}$ commonly exceeds C_{eb}.

The diffusion component of a collector-junction capacitance depends on a change in the collector junction width and consequently on the value of the charge with a change in voltage V_c. The expression for $C_{e\,dif}$ at $W/L_p \ll 1$ can be written in the form

$$C_{e\,dif} = I_e \frac{W}{D_p} \frac{\partial W}{\partial V_c} = I_e \sqrt{\frac{\varepsilon\varepsilon_0}{2qN}} \frac{W}{D_p \sqrt{V_c}} \tag{5.68}$$

A simple comparison of operating voltages across the emitter and collector gives us grounds to conclude that a substantially smaller change of the emitter voltage is required as compared to that of the collector voltage to vary the charge by the same value. This implies that the diffusion capacitance of the collector is smaller than that of the emitter. The diffusion capacitance of the emitter is usually hundreds and even more times higher than that of the collector.

Estimates show that in calculating the normal operating conditions of a transistor the barrier capacitance of the emitter may be disregarded as against its diffusion capacitance, while the barrier capacitance of the collector should be allowed for and its diffusion capacitance may be ignored.

5.6. FREQUENCY DEPENDENCE OF THE CURRENT GAIN

The alpha current gain remains essentially invariable in a broad range of frequency variation. However, as the frequency increases the alpha current gain begins to fall, limiting therewith the ability of the transistor to operate at still higher frequencies. The typical dependence of alpha on frequency f is given in Fig. 5.7. The frequency at which alpha has decreased to $1/\sqrt{2}$ (0.707) of its original value is called cutoff frequency (f_α).

One of the principal causes responsible for a decrease of alpha with changing frequency is the varying final time of transition of minority carriers across the base region. As is known, the minority carriers injected at one and the same moment reach the collector junction at different times, with the result that the shape of the output signal becomes distorted and its amplitude decreased. Besi-

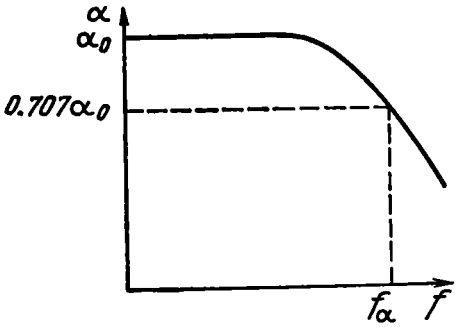


Fig. 5.7. Alpha *versus* frequency curve

des, there appears a phase delay between the changes of the emitter and collector currents. The distortion and phase delay of the collector current grow in proportion with the frequency of the signal and the transistor base width. As the frequency approaches its critical value the transistor ceases to amplify at all the applied high-frequency signal.

The alpha cutoff frequency must be inversely proportional to the time of transition of minority carriers through the base. This transit time is

$$t_{pb} = \frac{W}{v_{pb}} \tag{5.69}$$

where v_{pb} is the mean velocity of minority carriers in the base region.

If we assume that the transport of minority carriers is effected only by diffusion, which occurs in the case under discussion when the impurities are uniformly distributed in the base, then $v_{pb} = D_p/W$. Accordingly

$$t_{pb} = \frac{W^2}{D_p} \tag{5.70}$$

An accurate estimate gives the final expression for f_α

$$f_\alpha = \frac{1.22 D_p}{\pi W^2} \tag{5.71}$$

and

$$f_\alpha = \frac{A}{W^2}\ \text{Hz} \tag{5.72}$$

For germanium fused *p-n-p* and *n-p-n* transistors and silicon fused *p-n-p* and *n-p-n* transistors A = 17, 36, 5 and 12, respectively.

We can thus conclude that a higher limit of the alpha cutoff frequency is possible to achieve by cutting down the base width W.

5.7. TRANSISTOR NOISE

In a transistor, as in all conductors, there exist fluctuations of electric voltage (current) or the so-called electric noise. This noise must be taken into consideration when using a transistor as an amplifier of weak signals. A transistor is able to amplify a signal only if it proves discernible against the transistor's background noise. The intrinsic noise of a transistor may be described by a mean-square voltage (current) of noise or the noise factor, also called noise figure F_n, which is the ratio of the total noise power at the transistor output to the portion of noise power due to thermal (resistance) noise of the source of a useful signal.

One can differentiate several components of the intrinsic noise of a transistor.

1. *Thermal (resistance) noise* of a transistor, generated by the random thermal motion of charge carriers in the transistor bulk, mainly in the base region. The level of this noise is generally determined as a mean-square noise voltage

$$\overline{V}_{tn}^2 = 4kTr_b\Delta f \tag{5.73}$$

where Δf is the frequency interval within which the noise is generated. This type of noise is lowered by cutting down the value of r_b. The thermal noise is independent of frequency.

2. *Shot noise*, also called fluctuation or random noise, generated due to the random flow of charge carriers across the *p-n* junctions of a transistor. As a result of the static velocity-distribution of charge carriers, the number of carriers passing through a unit cross-sectional area in a unit time varies within certain limits, giving rise thereby to an alternating current component. The shot noise is frequency independent too.

For the emitter junction, the shot noise is given by

$$\overline{V}_{sne}^2 = 2qI_er_e\Delta f \tag{5.74}$$

and for the collector junction,

$$\overline{V}_{snc}^2 = 2qI_cr_c^2\Delta f \tag{5.75}$$

3. *Partition noise*, associated with the division of the emitter current in the base into a base and a collector current. Fluctuations in the distribution of minority carriers between the base and the collector produce noise characterized by a mean-square current amplitude equal to

$$\overline{i}_n^2 = 2qI_e(1-\alpha)\Delta f \tag{5.76}$$

This noise is significant when the base current is large, i.e., when alpha is small. This is normally the case at sufficiently high frequencies when alpha reduces markedly.

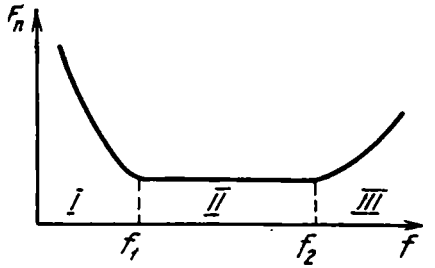


Fig. 5.8. Transistor noise factor as a function of frequency

4. *Low-frequency noise*, in inverse proportion to frequency. This type of noise prevails within the frequency range from fractions of a hertz to several kilohertz. The nature of the low-frequency noise has not been sufficiently studied so far. This noise is supposed to be associated with the transition of electrons to local trap centres and with their release from the traps. The trap centres may be disposed both within the volume and at the surface of a semiconductor. The trapping time of electrons in these centres may vary from fractions of milliseconds (fast traps) to several seconds (slow traps). The low-frequency noise level is appreciably dependent on the properties of the material used, type of its treatment and surface conditions. The plot of a noise figure-to-frequency curve is illustrated in Fig. 5.8.

Interval *I* (2 to 5 kHz) is noted for low-frequency noise, while intervals *II* and *III* feature shot and thermal noise, respectively. Beginning with the frequencies of approximately 0.1 f_α, the partition noise becomes predominant, whose value grows with increasing frequency due to a reduction in alpha.

It should be noted that much influence on the operation of a transistor also exerts the so-called leakage noise caused by the bridging currents present at the surface of *p-n* junctions due to surface contamination and inversion layers. This noise in transistors must be reduced to a minimum.

Review Questions

1. Draw a connection diagram for an *n-p-n* transistor and indicate the direction of current flow therein.
2. Plot curves for I_e, I_c, V_c and V_{out} of a *p-n-p* transistor amplifying a negative square wave.
3. Define the alpha current gain, transport efficiency β^* and injection efficiency γ, and describe their relationship.
4. Derive the expression describing alpha with and without regard for surface recombination.
5. Determine a change in the output current of a transistor if the emitter current has changed from 5 to 10 mA, alpha being equal to 0.96.

6. Determine a change in the emitter resistance with the emitter current changing from 5 to 10 mA.

7. In which of the two transistors with $\rho_b = 1$ ohm·cm or $\rho_b = 0.5$ ohm·cm does a punch-through of the emitter and the collector junction set in earlier?

8. Explain why the emitter diffusion capacitance is proportional to the current flow.

9. What width of the base must a silicon fused *p-n-p* transistor have for its alpha cutoff frequency to equal that of a germanium fused *n-p-n* transistor?

10. Plot and explain the curve of the frequency dependence of a transistor noise factor.

Chapter Six

TRANSISTOR AS A FOURPOLE

6.1. TRANSISTOR AS AN ACTIVE FOURPOLE

It is convenient to represent a transistor as a fourpole, i.e., a device having two input and two output terminals as is shown in Fig. 6.1 (also called quadripole, four-terminal network or two port).

Voltages (V_1, V_2) and currents (I_1, I_2) at the input and output of a fourpole are interrelated and characterize the operating point on the static characteristics of the device. It is sufficient to specify two out of the four quantities to determine graphically (by characteristics) or analitically the other two quantities. Therefore, taking two of them, for example, voltages, as independent variables (parameters), one can write down the characteristics of the fourpole in terms of functions of the parameters in the following form:

$$\left.\begin{aligned} I_1 &= I_1(V_1, V_2) \\ I_2 &= I_2(V_1, V_2) \end{aligned}\right\} \tag{6.1}$$

If sufficiently small variable signals are superimposed on the constant components of currents and voltages, their amplitude (or effective values) may be regarded as small increments of the constant components.

Considering that the current and voltage increments are rather small (small-signal mode) in comparison with the constant values of currents (I_{10} and I_{20}) and voltages, we can write

$$\left.\begin{aligned} I_{10} + \Delta i_1 &= I_{10} + \frac{\partial I_1}{\partial V_1}\Delta v_1 + \frac{\partial I_1}{\partial V_2}\Delta v_2 + \dots \\ I_{20} + \Delta i_2 &= I_{20} + \frac{\partial I_2}{\partial V_1}\Delta v_1 + \frac{\partial I_2}{\partial V_2}\Delta v_2 + \dots \end{aligned}\right\} \tag{6.2}$$

where small letter symbols designate variable components.

Expression (6.2) represents an expansion of fourpole currents into a Taylor series, which is convenient in determining the increments the dependent variables gain (currents I_1 and I_2 in our case).

Restricting ourselves only to the first terms of the expansion and identifying the derivatives in a new way, it is easy to obtain six systems of linear equations corresponding to six possible ways of

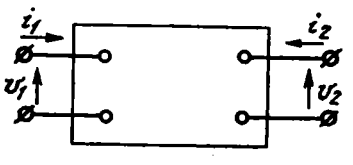


Fig. 6.1. Four-terminal network

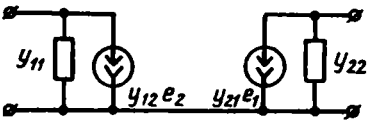


Fig. 6.2. Fourpole equivalent circuit as obtained from y parameters

selecting the parameters of a fourpole. The partial derivatives which serve as parameters (coefficients) in the equations can be either dimensionless quantities or can have the dimension of impedance or admittance. These parameters are normally identified with subscripts (indices) 11, 12, 21, 22. Subscript 1 denotes the values of quantities at the input (for example, v_1 —input voltage), subscript 2 stands for the values of quantities at the output (for example, i_2 — output current). Then, the parameter with subscript 11 is an input one, i.e., refers to the current and voltage at the input, the parameter with subscript 22 is an output one, that with subscript 21 describes the effect of the input on the output, i.e., the forward transfer, that with subscript 12 determines the influence of the output on the input, i.e., the reverse transfer.

Admittance Representation (y Parameters)

Quantities v_1 and v_2 are taken as independent variables in the fourpole equations:

$$\left.\begin{aligned} i_1 &= y_{11}v_1 + y_{12}v_2 \\ i_2 &= y_{21}v_1 + y_{22}v_2 \end{aligned}\right\} \qquad (6.3)$$

The y parameters in (6.3) have the dimension of admittances and are determined for an a.c. short-circuited input or output. Thus,

$$y_{11} = \frac{\partial i_1}{\partial v_1}\bigg|_{v_2=0}$$

is the input admittance for an a.c. short-circuited output;

$$y_{12} = \frac{\partial i_1}{\partial v_2}\bigg|_{v_1=0}$$

is the reverse transfer admittance, i.e., the quantity that describes the effect of changes in the output voltage on the short-circuit input current;

$$y_{21} = \frac{\partial i_2}{\partial v_1}\bigg|_{v_2=0}$$

is the forward transfer admittance, i.e., the quantity that determines the effect of the changing input voltage on the short-circuit output

current;

$$y_{22} = \frac{\partial i_2}{\partial v_2}\bigg|_{v_1=0}$$

is the output admittance for an a.c. short-circuited input.

The equivalent fourpole network is shown in Fig. 6.2.

Impedance Representation (z Parameters)

The equivalent fourpole circuit for the z representation is given in Fig. 6.3. Currents i_1 and i_2 are chosen as independent variables. The system of equations is given in the form

$$\left.\begin{aligned} v_1 &= z_{11}i_1 + z_{12}i_2 \\ v_2 &= z_{21}i_1 + z_{22}i_2 \end{aligned}\right\} \tag{6.4}$$

The parameters in this system of the fourpole representation have the dimension of impedances commonly measured under the open-circuit condition for a.c. Thus,

$$z_{11} = \frac{\partial v_1}{\partial i_1}\bigg|_{i_2=0}$$

is the input impedance for an a.c. open-circuited output;

$$z_{12} = \frac{\partial v_1}{\partial i_2}\bigg|_{i_1=0}$$

is the reverse transfer impedance, i.e., the quantity that describes the influence of changes in the output current on the open-circuit input voltage;

$$z_{21} = \frac{\partial v_2}{\partial i_1}\bigg|_{i_2=0}$$

is the forward transfer impedance, i.e., the quantity that characterizes the effect of changes in the input current on the open-circuit output voltage;

$$z_{22} = \frac{\partial v_2}{\partial i_2}\bigg|_{i_1=0}$$

is the output impedance for an a.c. open-circuited input.

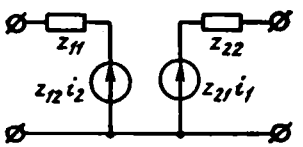


Fig. 6.3. Fourpole equivalent circuit using z parameters

h Representation

In this system of the fourpole representation i_1 and v_2 are independent variables. The fourpole equations for h parameters are

$$\left.\begin{aligned} v_1 &= h_{11}i_1 + h_{12}v_2 \\ i_2 &= h_{21}i_1 + h_{22}v_2 \end{aligned}\right\} \tag{6.5}$$

The equivalent fourpole circuit obtained from the h parameters is shown in Fig. 6.4.

The condition of measuring the h parameters most closely conforms to the operating conditions of the transistor, i.e., the short-circuited output and open-circuited input. The h parameters have the following form:

$$h_{11} = \frac{\partial v_1}{\partial i_1}\bigg|_{v_2=0}$$

is the input impedance for a short-circuited output;

$$h_{12} = \frac{\partial v_1}{\partial i_2}\bigg|_{i_1=0}$$

is the reverse voltage transfer ratio for an open-circuited output;

$$h_{21} = \frac{\partial i_2}{\partial i_1}\bigg|_{v_2=0}$$

is the forward current transfer ratio for a short-circuited output;

$$h_{22} = \frac{\partial i_2}{\partial v_2}\bigg|_{i_1=0}$$

is the output admittance for an open-circuited input.

There are three more systems which define g, a, and b parameters, with independent variables v_1 and i_2; i_2 and v_2; and i_1 and v_1, respectively, though these systems are not widely used.

Parameters of one system may be expressed in terms of the parameters of another system. Formulas for converting parameters are given in Table 6.1.

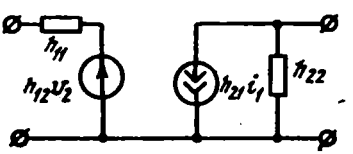


Fig. 6.4. Fourpole equivalent circuit using h parameters

Table 6.1

Conversion Relations Among the Systems of Fourpole Representations

Of system	To system		
	z	y	h
z	—	$y_{11}=\frac{z_{22}}{\Delta z}$; $y_{12}=\frac{-z_{12}}{\Delta z}$ $y_{21}=\frac{-z_{21}}{\Delta z}$; $y_{22}=\frac{z_{21}}{\Delta z}$	$h_{11}=\frac{\Delta z}{z_{22}}$; $h_{12}=\frac{z_{12}}{z_{22}}$ $h_{21}=\frac{-z_{21}}{z_{22}}$; $h_{22}=\frac{1}{z_{22}}$
y	$z_{11}=\frac{y_{22}}{\Delta y}$; $z_{12}=\frac{-y_{12}}{\Delta y}$ $z_{21}=\frac{-y_{21}}{\Delta y}$; $z_{22}=\frac{y}{\Delta y}$	—	$h_{11}=\frac{1}{y_{11}}$; $h_{12}=\frac{-y_{12}}{y_{11}}$ $h_{21}=\frac{y_{21}}{y_{11}}$; $h_{22}=\frac{\Delta y}{y_{11}}$
h	$z_{11}=\frac{\Delta h}{h_{22}}$; $z_{12}=\frac{h_{12}}{h_{22}}$ $z_{21}=\frac{-h_{21}}{h_{22}}$; $z_{22}=\frac{1}{h_{22}}$	$y_{11}=\frac{1}{h_{11}}$; $y_{12}=\frac{-h_{12}}{h_{11}}$ $y_{21}=\frac{h_{21}}{h_{11}}$; $y_{22}=\frac{\Delta h}{h_{11}}$	—

$$\Delta z=z_{11}z_{22}-z_{12}z_{21}=\frac{1}{\Delta y}=\frac{h_{11}}{h_{22}};\ \Delta y=y_{11}y_{22}-y_{12}y_{21};\ \Delta h=h_{11}h_{22}-h_{12}h_{21}=$$
$$=\frac{z_{11}}{z_{22}}=\frac{y_{22}}{y_{11}}$$

6.2. THREE BASIC TRANSISTOR CONFIGURATIONS

There are three ways of transistor connection depending on which electrode of the transistor is common. Fig. 6.5 illustrates three connection diagrams for a transistor. In analyzing the connection dia-

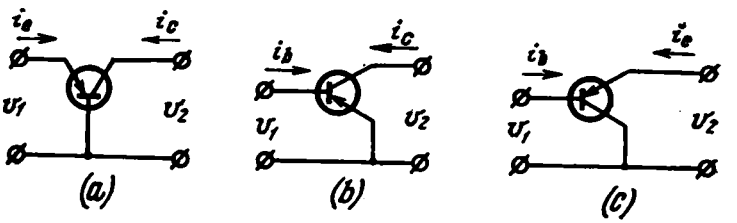


Fig. 6.5. Basic transistor configurations
(*a*) common base; (*b*) common emitter; (*c*) common collector

grams one should bear in mind that the physical processes taking place in a transistor do not depend on which of the electrodes is chosen to be common.

Common-Base Connection

The above discussions in Chapters Five and Six relate only to the transistor in a common-base connection. As has been mentioned in Sec. 5.2, the current gain α (now it will be denoted by h_{21b}) in the common-base circuit is, as a rule, less than unity. But owing to a low input impedance of the transistor as compared to its output impedance, the circuit is able to adequately amplify the voltage and power of a signal, which makes this type of connection most popular. Frequency properties of a common-base transistor have been discussed in detail in Sec. 5.6.

Common-Emitter Connection

This is the most extensively used type of transistor connection which allows for the largest power and current amplification. Indeed, the current equation for a common base transistor, as is evident from (5.28) and (5.29). can be written in the form

$$i_e = i_c + i_b = i_e h_{21b} + i_b$$

It is obvious that the common-base connection does not provide for current amplification because the control current here is the emitter current and $h_{21b} < 1$. In the common-emitter connection, the control current is the current of the base

$$i_b = i_e (1 - h_{21b})$$

The current gain (the current-transfer ratio) for the common-emitter circuit is

$$h_{21e} = \frac{i_c}{i_b} = \frac{h_{21b}}{1 - h_{21b}} \tag{6.6}$$

Formula (6.6) shows that a substantial rise in h_{21e} corresponds to an insignificant rise in h_{21b}. For example, with a change of h_{21b} from 0.96 to 0.98, h_{21e} grows from 24 to 49, i.e., more than two times.

The current-gain cutoff frequency f_{h21e} in the common-emitter transistor is lower than the f_{h21b} in the common-base transistor and is related to the latter quantity by

$$f_{h21e} \approx \frac{f_{h21b}}{h_{21e0}(1+m)} \tag{6.7}$$

where m is the coefficient dependent on the design of a transistor and in general ranges from 0.6 to 0.8, and h_{21e0} is the value of h_{21e}

at a low frequency (up to 1 kHz). The notion of a maximum current-gain frequency f_{max} is rather often introduced in practice. At the frequency f_{max} the current gain $|h_{21e}|$ becomes equal to 1. The quantity f_{max} is h_{21e} times as high as f_{h21e}. It should be noted that at appreciably high frequencies ($f \geqslant 3f_{h21e}$) the product $|h_{21e}|f = f_{max} = \text{const}$. This enables one to determine f_{max} by measuring $|h_{21e}|$ at a considerably lower frequency as compared to f_{max}.

Common-Collector Connection

In this type of circuit, as in the common-emitter circuit, the base current serves as a control current, but the output current is drawn from the emitter.

The current gain in the common-collector circuit is

$$h_{21c} = \frac{i_e}{i_b} = \frac{1}{1 - h_{21b}} \tag{6.8}$$

Keeping in mind that h_{21b} in transistors is generally close to unity, it follows from the comparison of (6.8) with (6.6) that the current gain of the common-emitter circuit differs but rather insignificantly from that of the common-collector circuit. As a consequence, the frequency properties of these types of circuits also differ but little.

The input voltage V_{bc} in the common-collector connection is lower than V_{ec} by a small value of V_{eb}, while the values of input current I_b and output current I_e differ by many times ($I_b \ll I_e$). From this it follows that the common-collector circuit has a high input and a low output impedance, this being an important feature of the common-collector circuit.

The common-collector transistors are more rarely used than common-emitter or common-base transistors.

Transistor Equivalent Circuits

To simulate the physical phenomena that occur in a transistor by representing it in the form of a fourpole is not always convenient and the fourpole is often replaced by a T equivalent circuit (Fig. 6.6).

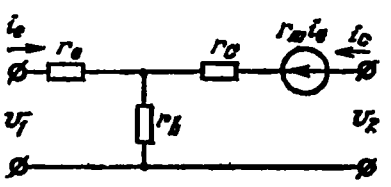


Fig. 6.6. T equivalent circuit for the transistor using a voltage generator

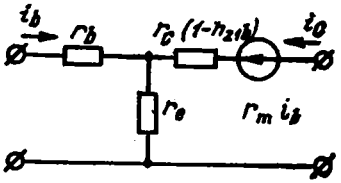


Fig. 6.7. Voltage generator equivalent circuit for the transistor in a common-emitter connection

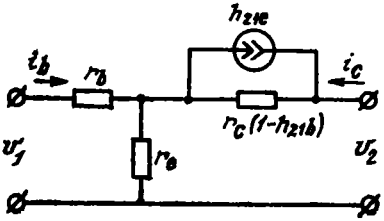


Fig. 6.8. Current generator equivalent circuit for the transistor in a common-emitter connection

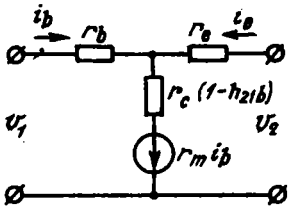


Fig. 6.9. Voltage generator equivalent circuit for the transistor in a common-collector connection

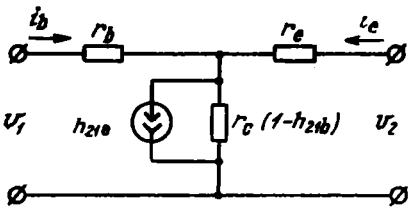


Fig. 6.10. Current generator equivalent circuit for the transistor in a common-collector connection

By comparing the equations for the z parameters of the fourpole representation

$$\left.\begin{aligned} v_1 &= r_{11}i_1 + r_{12}i_2 \\ v_2 &= r_{21}i_1 + r_{22}i_2 \end{aligned}\right\} \qquad (6.9$$

with the equations for the equivalent circuit of Fig. 6.6

$$\left.\begin{aligned} v_1 &= (r_e + r_b)\, i_e + r_b i_c \\ v_2 &= (r_b + r_m)\, i_e + (r_b + r_c)\, i_c \end{aligned}\right\} \qquad (6.10)$$

Table 6.2

Parameters of fourpole	Parameters of equivalent circuit		
	common-base connection	common-emitter connection	common-collector connection
r_{11}	r_e+r_b	r_e+r_b	r_c
r_{12}	r_b	r_e	$r_c(1-h_{21b})$
r_{21}	$r_c h_{21b}$	$-r_c h_{21b}$	r_c
r_{22}	r_c	$r_c(1-h_{21b})$	$r_c(1-h_{21b})$
y_{11}	$\frac{1}{r_e+r_b(1-h_{21b})}$	$\frac{1-h_{21b}}{r_e+r_b(1-h_{21b})}$	$\frac{1-h_{21b}}{r_e+r_b(1-h_{21b})}$
y_{12}	$-\frac{r_c}{r_c[r_e+r_b(1-h_{21b})]}$	$-\frac{r_e}{r_c[r_e+r_b(1-h_{21b})]}$	$-\frac{1-h_{21b}}{r_e+r_b(1-h_{21b})}$
y_{21}	$-\frac{h_{21b}}{r_e+r_b(1-h_{21b})}$	$\frac{h_{21b}}{r_e+r_b(1-h_{21b})}$	$-\frac{1}{r_e+r_b(1-h_{21b})}$
y_{22}	$\frac{r_e+r_b}{r_c[r_e+r_b(1-h_{21b})]}$	$\frac{r_e+r_b}{r_c[r_e+r_b(1-h_{21b})]}$	$\frac{1-h_{21b}}{r_e+r_b(1-h_{21b})}$
h_{11}	$r_e+r_b(1-h_{21e})$	$\frac{r_e}{1-h_{21b}}+r_b$	$\frac{r_e}{1-h_{21b}}+r_b$
h_{12}	$\frac{r_b}{r_c}$	$\frac{r_e}{r_c(1-h_{21b})}$	1
h_{21}	$-h_{21b}=\frac{2m}{r_c}$	$\frac{h_{21b}}{1-h_{21b}}$	$-\frac{h_{21b}}{1-h_{21b}}$
h_{22}	$\frac{1}{r_c}$	$\frac{1}{r_c(1-h_{21b})}$	$\frac{1}{r_c(1-h_{21b})}$

it is easy to obtain the agreement between the parameters

$$\left.\begin{aligned} r_{11} &= r_e + r_b \\ r_{12} &= r_b \\ r_{21} &= r_b + r_m \\ r_{22} &= r_b + r_c \end{aligned}\right\} \qquad (6.11)$$

$$\left.\begin{aligned} r_e &= r_{11} - r_{12} \\ r_b &= r_{12} \\ r_c &= r_{22} = r_{12} \\ r_m &= r_{21} - r_{12} \end{aligned}\right\} \qquad (6.12)$$

The amplification ability of a transistor is reflected in the equivalent circuit by a dependent voltage generator with internal resistance r_m.

It is also possible to construct a transistor T equivalent circuit with a current generator. Such a circuit was considered in Chapter Five (see Fig. 5.4).

Analogous equivalent circuits for transistors in a common-emitter and a common-collector configuration are given in Figs. 6.7 through 6.10. The formulas relating the h, y, and z parameters of fourpoles to the physical parameters of the equivalent circuits of transistors are listed in Table 6.2.

6.3. STATIC CHARACTERISTICS OF THE TRANSISTOR

As has been shown in Sec. 6.1, the transistor characteristics may be represented by the six systems of parameters which interrelate by way of functions the four quantities, namely, currents I_1 and I_2, and voltages V_1 and V_2.

Adopted as the main sets of characteristics are those which interrelate the current and voltage at the input and output, respectively called a set of input and a set of output characteristics. Correspondingly, the output current and voltage are the parameters for input characteristics, and the input current and voltage are the parameters for output characteristics.

The sets of characteristics which relate the currents or voltages at the input to the currents or voltages at the output are referred to as the forward transfer characteristics. They are not independent, as are not the characteristics relating the currents or voltages at the output to the currents or voltages at the input (reverse transfer characteristics), and may be obtained from the main characteristics.

We now consider static current-voltage characteristics of transistors in different connections.

The input characteristics of a transistor in a common-base connection (Fig. 6.11*a*) bear a resemblance to the diode characteristics

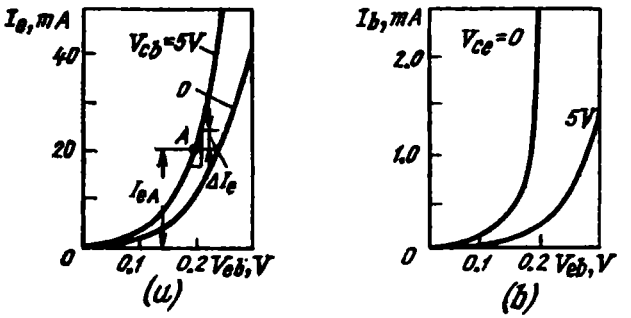


Fig. 6.11. Static input characteristics of the transistor in a common-base connection (*a*), and in a common-emitter connection (*b*)

(see Fig. 2.7) and correspond to them at $V_{cb} = 0$. With increasing V_{cb} the thickness of the space-charge layer increases too and the base region becomes correspondingly thinner, which raises the emitter current at the same value of V_{eb}. At a specified value of the emitter current, an increase in V_{cb} is attended by a decrease in V_{eb}.

Common-emitter input curves (Fig. 6.11*b*) and common-base input curves are very much alike. However an increase of the voltage across the collector of the transistor in the common-emitter connection, as contrasted to the common-base connection, causes the base current to diminish at the same value of V_{eb} (the shift of the curve to the right). This is attributed to a decrease in the base width with rising collector voltage and, according to (5.30), to an increase in the current gain h_{21b}. This in turn leads to a reduction in the portion of the base current as compared to the emitter current increment producing an additional voltage drop across the emitter junction.

The output characteristics of a transistor in a common-base connection are shown in Fig. 6.12*a*, which at $I_e = 0$ correspond to the reverse curves of a diode. The collector current saturation occurs here already at small values of V_c. An extremely gentle slope of the curves at $I_e > 0$ is due to a weak effect of the collector voltage on the current because the injected carriers are distributed over the base mainly by diffusion.

For the same reason almost all the current of minority carriers reaches the collector junction at $V_{cb} = 0$, and for the collector current to reduce to zero, it is necessary to apply to the collector a small voltage of opposite sign, causing an equal flow of carriers from the collector junction. The output characteristics of a transistor in a common-emitter connection (Fig. 6.12*b*) pass through the origin and do not cover the region of the collector voltage of opposite sign. The dependence of I_c on V_c is more pronounced here than in the case of

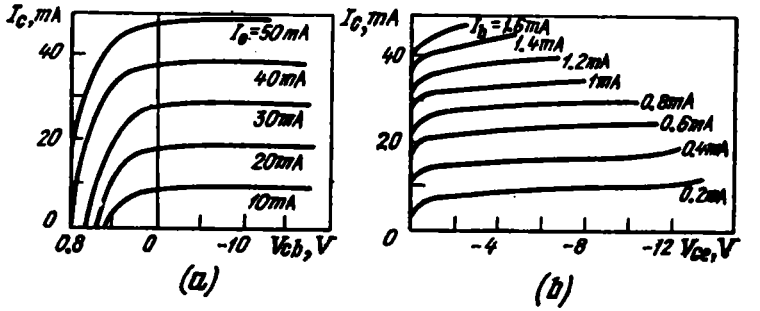


Fig. 6.12. Static output characteristics of the transistor in a common-base connection (*a*), and in a common-emitter connection (*b*)

the common-base connection because the values of I_b are related to the values of I_c which sharply grow with increasing V_c, this being accompanied, as is apparent from formulas (5.30) and (6.6), by a decrease in the effective base width and an increase in the current gain. Since voltages V_{ce} and V_{be} are connected in opposition the collector junction derives the voltage from the difference between V_{ce} and V_{be}. Hence, at $|V_c| < |V_b|$ the collector junction is rendered conductive in the forward direction. Small variations in V_c are sufficient to obtain large variations in I_c. At $|V_c| > |V_b|$ the collector junction becomes reverse-biased and variations in V_c affect but little the value of I_c.

One can determine h parameters from the static characteristics: h_{11} and h_{12} from the input curves and h_{21} and h_{22} from the output curves. For example, augmenting or decreasing the emitter current by $\pm\Delta I_e$ near the selected point A (Fig. 6.11*a*), we find increments or decrements $\pm\Delta V_{eb}$ on the current-voltage curve and then easily obtain the voltage-current increment (decrement) relationship

$$h_{11b} = \frac{2\Delta V_{eb}}{2\Delta I_e} = \frac{0.02}{8\times 10^{-3}} = 2.5 \text{ (ohms)}$$

In a similar manner we can determine h_{11} parameters for the common-emitter connection (Fig. 6.11*b*) and other h parameters in any transistor connection diagrams.

6.4. DEPENDENCE OF TRANSISTOR MAIN PARAMETERS ON OPERATING CONDITIONS AND TEMPERATURE

Transistor parameters greatly depend on the operating conditions. Therefore, in order to make the best use of the transistor, its mode of operation should be so selected on the basis of the function of the transistor and the properties of its circuit that it can be conformable

to the optimum transistor parameters, important for the purpose. The dependence of the parameters of a transistor T equivalent circuit on current and voltage is illustrated in Fig. 6.13.

The current transfer ratio-versus-emitter current curve has a maximum corresponding to the optimum current density. Though formula (5.30) does not show the direct relationship between h_{21e} and I_e, in real conditions the increase in I_e at small values of the current gives rise to an internal field in the base and causes a certain decrease in recombination and consequently an increase in h_{21e} (Fig. 6.13*a*). Further rise in the current density increases recombination losses in proportion to the number of minority carriers and decreases the emitter effectiveness owing to growing conductivity of the base region, which subsequently results in lowering of h_{21e}. The nature of these phenomena is rather complicated and will not therefore be considered here in detail. The geometry of transistor regions and the pattern of the impurity distribution play an important role in the development of these phenomena.

As is seen from Fig. 6.13*b*, h_{21e} rises with increasing voltages V_c. This rise at small voltages is attributed to decreasing recombination in view of the narrowing base region. At higher voltages the avalanche multiplication in the collector *p-n* junction becomes a determining factor in the sharp increase of h_{21e}.

The *emitter resistance*, as is obvious from (5.33), falls with growing current (Fig. 6.13*a*) and does not depend on the voltage across the collector junction (Fig. 6.13*b*).

In conformity with (5.42), the *collector resistance* diminishes with increasing I_e (Fig. 6.13*a*) and rises as soon as V_c grows (Fig. 6.13*b*). However, at values of V_c approaching the breakdown voltage of the collector junction, the collector resistance begins to fall due to the avalanche multiplicati n of charge carriers in the collector *p-n* junction

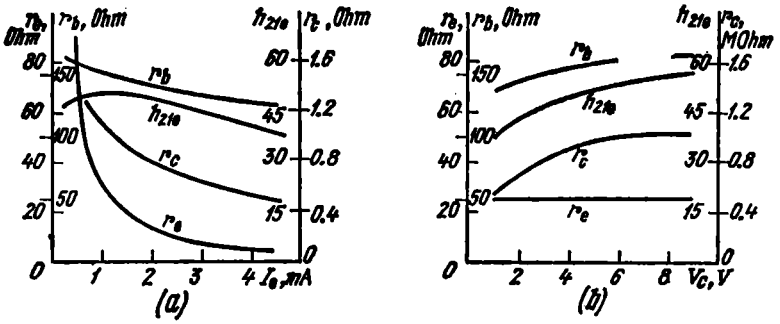


Fig. 6.13. Main transistor parameters as a function of emitter current (*a*), and collector voltage (*b*)

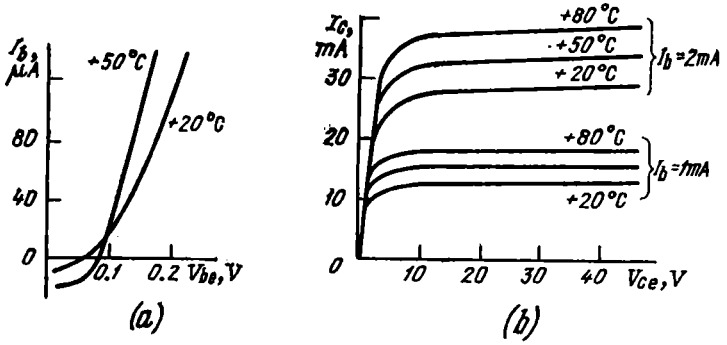


Fig. 6.14. Input (*a*) and output (*b*) characteristics of a transistor at elevated temperatures

The decline in the *base resistance* r_b with growing I_e (Fig. 6.13*a*) is attributed to the increase in the density of minority and majority carriers. The rise in r_b with increased V_c (Fig. 6.13*b*) is associated with a change in the base geometry (a decrease in the cross-sectional area of the base in a real junction transistor) due to the expansion of the space-charge layer near the collector *p-n* junction.

The main relationships determining the transistor parameters have been derived earlier in Chapter Five. As a rule, almost all the expressions contain the terms with temperature T. This is quite natural since the density of minority carriers is also determined by temperature, apart from other factors.

The *collector reverse current* I_{c0} is particularly dependent on temperature. As is clear from (5.54), this dependence may be expressed in the form

$$I_{c02} = I_{c01} e^{A(T_2 - T_1)}$$

where $A = 0.05$-0.09 for Ge; $A = 0.07$-0.13 for Si; and I_{c01} and I_{c02} are related to T_1 and T_2, respectively. A rather simple calculation shows that with rising temperature the current I_{c0} doubles as against its initial value every 6-8°C for silicon transistors and 9-11°C for germanium transistors.

Such a sharp rise in I_{c0} with increasing temperature is obviously objectionable. If I_{c0} becomes comparable with the working collector current, the transistor ceases to operate. At normal and increased temperatures, the surface condition has a significant effect on the value of I_{c0} and its temperature dependence.

The *static characteristics* of a transistor also undergo substantial changes with temperature. A heavy influence of temperature on the value of I_{c0} and consequently a decrease in the emitter input impedance, leading to the current rise in the emitter at its constant voltage, cause a shift of static characteristics (Fig. 6.14).

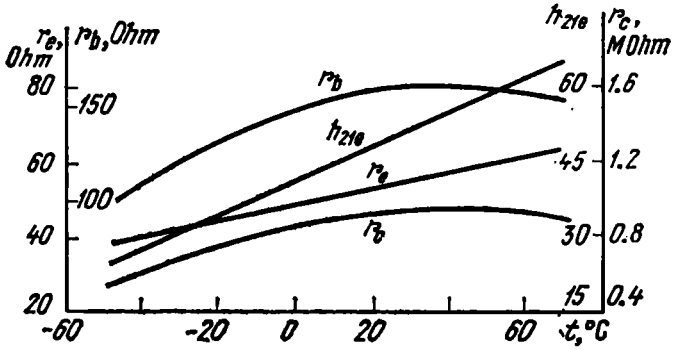


Fig. 6.15. Temperature dependence of main transistor parameters

As has been described in Sec. 5.2, the *current gain* (current transfer ratio) is determined by many factors, each in its own way being temperature dependent. Thus, temperature tells on the diffusion length and diffusion factor of minority carriers, surface recombination velocity, and the diffusion length of majority carriers. All these parameters enter into the expression for the current gain h_{21}. Of much importance for the temperature dependence of h_{21} are the initial stages in the production process of a transistor, its surface preparation, and thermal treatment. The h_{21} varies differently with temperature in different transistor types, though in most cases it rises with increasing temperature (Fig. 6.15).

According to formula (5.33), the *emitter resistance* rises with temperature. The temperature dependence of the collector resistance, shown in Fig. 6.15, is of a complicated nature.

The *current-gain cutoff frequency* falls with rising temperature since the diffusion factor $D = \mu \frac{kT}{q}$ diminishes at elevated temperatures and because $\mu \sim T^{-3/2}$ in the range of working temperatures for germanium and silicon transistors.

6.5. PULSE (SWITCHING) RESPONSE OF THE TRANSISTOR

The transistor behaviour in the pulse (switching) mode underlies the function of pulse circuits. By a switch is meant a circuit whose main function is to open and close the supply and load circuits on applying the control input signals.

By analogy with an electric (contact) switch, the quality of a transistor switch is determined primarily by the minimum voltage drop across it in the closed (ON) condition, the minimum current in the open (OFF) condition and also the speed of switching from one state to another.

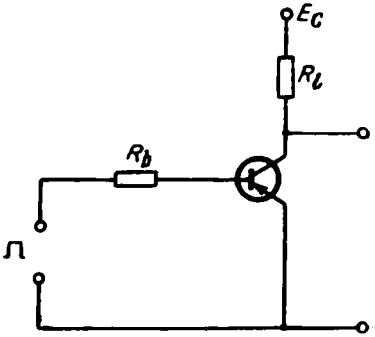


Fig. 6.16. Circuit of a transistor switch

Despite the simplicity of functions performed by the transistor switch, there are a variety of its circuit arrangements, the simplest one being illustrated in Fig. 6.16.

Load R_l is connected into a collector circuit, and control pulses are supplied to the transistor input via resistor R_b. Depending on the combination of values and polarities of the applied voltages, the quiescent or operating point of the switching transistor may lie in one of the four regions of operation: cutoff region *I*, active region *II*, saturation region *III* and avalanche multiplication region *IV* (Fig. 6.17). In region *I* both the emitter and the collector junction are reverse-biased (the so-called cutoff mode). In region *II* the active or normal mode of amplification sets in: the emitter junction injects minority carriers into the base (forward bias), while the collector junction is closed (reverse bias). In region *III* both the emitter and the collector junction are forward-biased and inject carriers of current into the base. Region *IV* is the area of avalanche multiplication (the current transfer ratio is equal to or is more than unity).

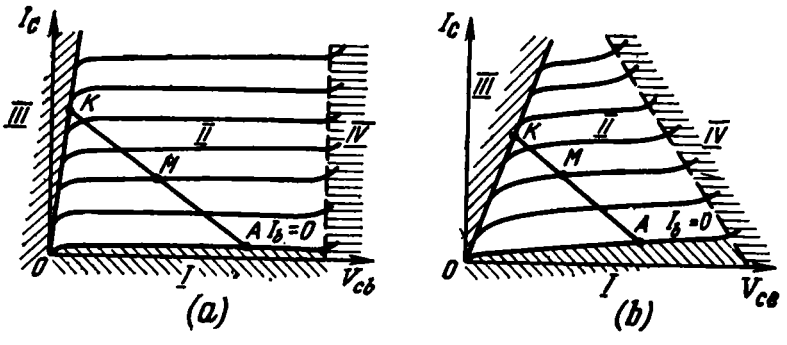


Fig. 6.17. Regions of operation of a transistor

(*a*) in common-base connection; (*b*) in common-emitter connection; *I*—cutoff region *II*—active region; *III*—saturation region; *IV*—avalanche multiplication region

If we plot the line of load R_l on the current-voltage curves and if $V_b = 0$ ($I_b = 0$) we find that the initial current I_{ce0} flows through the collector and the quiescent point is point A. The value of collector current I_{ce0} differs from that of collector reverse current I_{c0} because the latter is measured at $I_{eb} = 0$, and this condition is not fulfilled on the plot of Fig. 6.17. Owing to a small I_{ce0} the collector may be regarded to be at full voltage E_c. This condition is referred to as the cutoff state.

If we now increase V_b and consequently the base current, the quiescent point A becomes shifted along the line R_l towards point K. At a particular value of base current I_b the quiescent point coincides with point M. In this case the collector current is defined by quantities E_c and R_l since the voltage drop across the transistor may be neglected:

$$I_c = \frac{E_c}{R_l}$$

There is still a small voltage drop across the collector which is called saturation voltage. This condition of the switch is commonly identified as the closed state of the switch or the saturated and forward bias state of the transistor.

Further increase of the base current has no effect on the position of the quiescent point, but in actual circuits I_b is always raised to exceed I_{b1} at the boundary of the saturation region lest small variations in the supply voltage or thermal drift of parameters should shift point K along the load line R_l.

The ratio of the operating base current I_b to the base current I_{b1} at the boundary of the saturation region is called the saturation coefficient of a transistor

$$s_s = \frac{I_b}{I_{b1}} \tag{6.13}$$

If a cutoff (blocking) signal is applied to the transistor emitter, the negative voltage in the collector is not able to set up instantaneously and for some time the collector current remains higher than the current I_{ce0} under the influence of the charge being accumulated. The charge starts decaying due to the movement of holes from the base into the collector and the emitter junction. It is not until the hole density reaches an equilibrium value that the quiescent point becomes set in at A.

The transition of the quiescent point from the cutoff mode to the saturated mode does not occur instantaneously either and lasts for some time until the nonequilibrium carrier density reaches the magnitude corresponding to the saturated mode. This transit time is determined by the carrier-diffusion time constant.

The emitter and collector current pulses formed in the course of transition from the cutoff mode to the saturated mode and vice versa are shown in Fig. 6.18.

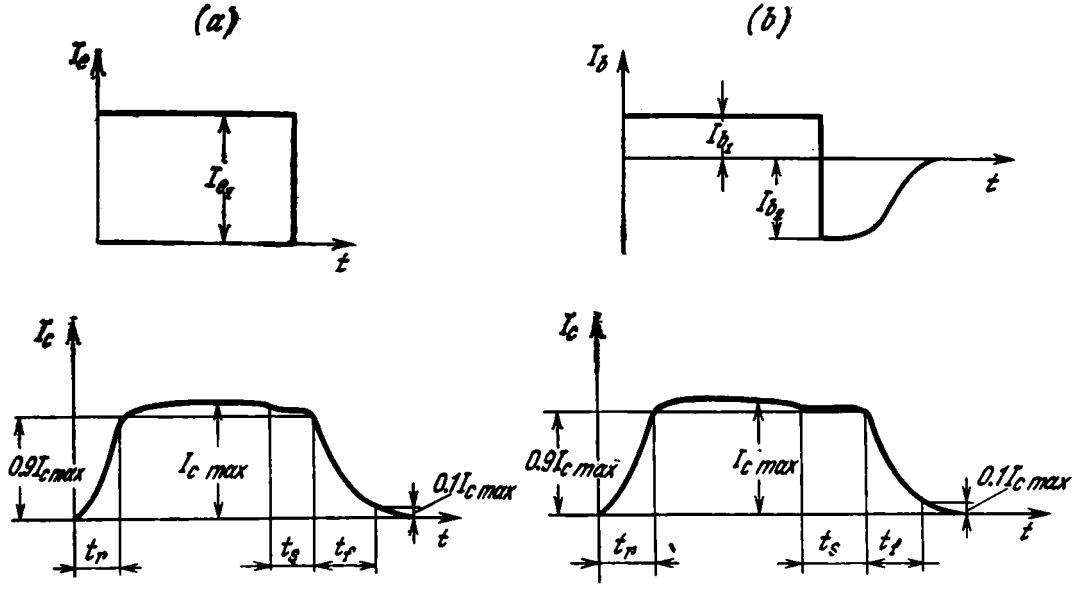


Fig. 6.18. Pulse waveforms at the input and the output of the transistor operating as a switch in a common-base connection (*a*), and in a common-emitter connection (*b*)

These processes determine the time characteristics of a transistor switch, which at specified parameters of the control pulses are defined by the three quantities: rise time (turn-on time) t_r, storage time t_s, and decay or fall time t_f.

By the *rise time* is meant the time the transistor takes to leave the cutoff state for the saturated mode.

The *storage time* t_s and the *fall time* t_f make up the *turn-off time* during which the transistor remains forward-biased after the forward-bias pulse has been cut off (Fig. 6.18).

The time t_r is determined by the time the collector current I_c takes to rise to 0.9 of its final value $I_{c\,max}$, and the time t_f is found from the time required for I_c to drop to 0.1 of $I_{c\,max}$, this being reflected in the expressions for determining t_r and t_f for the three types of transistor connection (Table 6.3).

Table 6.3

Expressions for Determining t_r and t_f of the Transistor

Transistor connection	t_r	t_f
Common base	$\frac{1}{2\pi f_{h21b}} \ln \frac{h_{21b}I_{e1}}{h_{21b}I_{e1}-0.9\,I_{c\,max}}$	$\frac{1}{2\pi f_{h21b}} \ln \frac{I_{c\,max}+h_{21b}I_{e2}}{0.1I_{c\,max}+h_{21b}I_{e2}}$
Common emitter	$\frac{1}{2\pi f_{h21e}} \ln \frac{h_{21e}I_{b1}}{h_{21e}I_{b1}-0.9\,I_{c\,max}}$	$\frac{1}{2\pi f_{h21e}} \ln \frac{0.9I_{c\,max}-h_{21e}I_{b2}}{0.1I_{c\,max}-h_{21e}I_{b2}}$
Common collector	$\frac{1}{2\pi f_{h21e}} \ln \frac{h_{21b}h_{21e}I_{b1}}{h_{21e}I_{b1}-0.9\,I_{e1}}$	$\frac{1}{2\pi f_{h21e}} \ln \frac{h_{21b}I_{e1}-h_{21e}I_{b2}}{0.1h_{21b}I_{e1}-h_{21e}I_{b2}}$

Note. I_1 is the forward-bias current pulse, and I_2 is the reverse-bias current pulse.

These expressions show that the better the frequency properties of a transistor, the lower the values of t_r and t_f. Let us also note that for the storage time t_s to be cut down, the lifetimes of minority carriers must be substantially small, which is commonly achieved by applying special methods during the production process. Of these methods the most widespread one mentioned in Chapter Four is the introduction in the semiconductor crystal of impurities, as gold or nickel, which reduce the effective lifetimes of minority carriers.

The time t_s in germanium low-frequency switching transistors ranges, on average, from 0.3 to 0.5 μs, and in silicon planar transistors, from 20 to 50 ns.

The saturated mode of operation of a forward-biased transistor is noted for a residual voltage (saturation voltage) V_{ces} and V_{bes}

between the collector and emitter and between the base and emitter, respectively.

For germanium transistors, $V_{bes} = 0.4$ to 0.5 V and $V_{ces} = 0.1$ to 0.2 V and for silicon transistors, $V_{bes} = 0.9$ to 1.1 V and $V_{ces} = 0.3$ to 0.4 V.

6.8. MAXIMUM PERMISSIBLE OPERATING CONDITIONS OF THE TRANSISTOR

In order to make the most use of a transistor, it is necessary to know the maximum permissible values of its main parameters.

The *maximum permissible parameters* are the upper limiting values the increase of which may impair the characteristics of a transistor or cause its failure. These parameters include the maximum permissible values of power, currents, voltages, mechanical stresses and temperature. The use of transistors under maximum permissible conditions is not allowed by specifications.

The *minimum permissible parameters* are the values the decrease of which may degrade the operation of the transistor in a circuit. These parameters are the minimum values of currents and voltages at which a transistor is still capable of amplifying a signal or the minimum temperature corresponding to the permissible temperature-dependent variations of the parameters.

Consider the most important maximum permissible parameters which define the conditions of transistor performance in various applications.

The *maximum reverse voltages* of a collector and an emitter junction have been discussed in detail in the pertinent Sections dealing with the properties of *p-n* junctions (Sec. 2.6) and transistors (Sec. 5.4). The maximum permissible voltage $V_{ce\,max}$ or V_{α} of the transistor in a common-emitter connection is also specified alongside the values of $V_{cb\,max}$ and $V_{be\,max}$. The allowable resistance in the emitter-base circuit is stipulated too in this type of transistor configuration because $V_{ce\,max}$ depends on R_{be} (see Sec. 5.4).

The *maximum currents* of a transistor are limited due to a number of reasons. For one thing, the rise of a maximum permissible current $I_{c\,max}$ may result in local overheating, for another, the rise of a specified invariable (or pulse) value of the maximum permissible current may cause an undesirable reduction in the current transfer ratio whose dependence on the emitter current has been discussed earlier.

The *maximum dissipated power* should be also given along with the maximum values of the collector voltage and current, since the dissipated power is mainly determined by the conditions of removal of the Joule heat released by an operating transistor and not by the dielectric strength margin.

The maximum permissible power of a transistor depends on the three temperature parameters: the maximum temperature of the junction, thermal resistance and heat capacity of the case. Besides, the ambient temperature and conditions of heat withdrawal have an appreciable effect on the maximum permissible transistor power.

Maximum temperature $t_{j\,max}$ of a junction is determined by the physical properties of the semiconductor material used in the manufacture of a transistor. In a semiconductor with a larger width of the band gap, $t_{j\,max}$ proves higher than in a semiconductor with a smaller band gap width because at higher temperatures the density of the charge carriers generated by heat becomes comparable with the impurity concentration. As has been mentioned above, for silicon devices $t_{j\,max} = 125$ to 175°C, for germanium devices $t_{j\,max} = 70$ to 85°C. It is natural that with increasing resistivity of one and the same material $t_{j\,max}$ decreases.

Thermal resistance R_t defines the temperature rise of a junction in degrees Celsius per unit dissipated power

$$t_j = PR_t + t_{am}$$

where P is the dissipated power, and t_{am} is the ambient temperature in °C. Hence,

$$R_t = \frac{t_j - t_{am}}{P} \tag{6.14}$$

Thermal resistance is sometimes represented as a sum of thermal resistance R_{tjc} between the junction and the case and thermal resistance $R_{t\,cen}$ between the case and environment. The first component depends on the design of the case and the method of bounding the crystal to the case, and the second component, on the case design, emission of heat from the case, environmental conditions, and additional means of heat removal, such as heat sinks, airflow, etc.

Thermal capacity C_t (W·s/°C) defines the amount of energy expended on heating a transistor by 1°C. The value of C_t determines the transient process of heating or cooling a transistor when the latter is turned on or off, and also when the ambient temperature changes abruptly.

The dissipated power should not exceed its maximum permissible value or else the transistor may be overheated and its parameters impaired which may even make it inoperative.

As the emitter voltage is very small the energy given out by the emitter junction when the transistor operates as a pulse amplifier is insignificant and may be disregarded, and all the power may be by right considered to be produced only by the collector junction. This is why the removal of heat from the collector junction is of primary concern. For this purpose the collector region of the transistor structure is usually made fastened to the case or a special heat exchan-

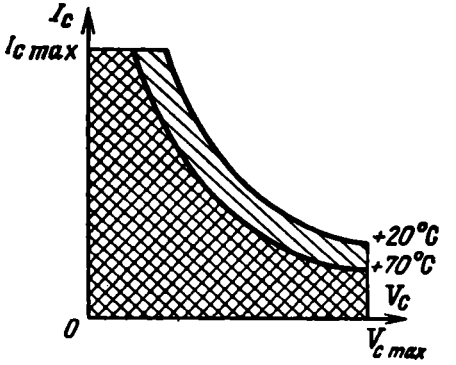


Fig. 6.19. Region of permissible collector currents and voltages for a transistor operating as a pulse amplifier

ger. This problem will be discussed in more detail in Chapter Seven.

If the transistor operates as a switch, account should be taken of the power dissipated by the emitter and the base during switching.

The choice of the quiescent point of the transistor operating in the active region presupposes the fulfilment of the following three conditions: $I_c < I_{c\,max}$, $V_c < V_{c\,max}$, and $P_c < P_{c\,max}$. These limiting conditions are illustrated graphically in Fig. 6.19. The quiescent point must lie inside the area bounded by the hyperbola $I_cV_c = P_{c\,max}$ at given ambient temperatures and maximum permissible values of current $I_{c\,max}$ and voltage $V_{c\,max}$.

The reliability of a transistor depends to a large measure on service conditions and a chosen mode of operation. Therefore, to provide for a prolonged fail-safe operation of the device as a whole, certain safety factors are made use of to allow the transistors to operate in somewhat alleviated conditions.

Maximum and minimum temperatures limit the temperature range within which the transistor can normally perform its functions. The limits of this range are determined from the allowable temperature-dependent variations of the parameters, the difference in the linear expansion coefficients of mating parts, properties of solders and plastic materials used, etc.

Review Questions

1. Draw three basic transistor connection diagrams.
2. Plot and explain typical input and output curves for the transistor in a common-base connection.
3. How are the h parameters determined from the static characteristics?
4. Plot and explain characteristic curves for h_{21}, f_{h21}, r_e, r_{ce}, r_b, C_c, and C_e as functions of transistor modes of operation.
5. Plot and explain the curves of the temperature dependence of h_{21}, f_{h21}, r_e, and I_{c0}.
6. Describe the operation of a transistor in an active and a saturation region.
7. What is the value of the maximum permissible dissipated power of a germanium transistor with $R_t = 0.3$°C/mW at +60°C?

Chapter Seven

TRANSISTOR TYPES AND APPLICATIONS

7.1. CLASSIFICATION OF TRANSISTORS

Transistors are one of the most widespread types of semiconductor devices. They may be classified by the material they are made of, mechanism of motion of minority carriers in the base, their function, electric and working parameters, design features and techniques of manufacture.

In the early period of their development, transistors were made only from germanium. In the subsequent years after science and technology have overcome difficulties associated with silicon purification, silicon transistors were devised that are not only comparable but rather superior in a number of parameters to germanium transistors.

The larger band gap width of silicon (1.11 eV), as compared to that of germanium (0.67 eV), results in better characteristics of silicon transistors. The collector thermal current in silicon transistors is by one-three orders of magnitude smaller than that of germanium devices. Silicon devices can operate at higher working voltages than germanium transistors.

The low factors of diffusion of boron, phosphorus, antimony and other alloying impurities into silicon dioxide and the attendant masking ability of the oxide displayed in the process of diffusion into silicon have provided the possibility for a high-precision local doping of silicon. This in turn has offered certain advantages in the manufacture of silicon devices and led to the creation of radically new promising methods of production.

Despite the mentioned merits of the silicon transistors, they supersede but rather slowly germanium devices in some branches of electronic industry, and to date germanium transistors are still the cheapest and enjoy wide use.

Intensive research into the creation of transistors from new semiconductor materials, such as gallium arsenide and indium phosphide, have been conducted in the last few years though these types of transistors have not yet passed the stage of laboratory investigations.

By the mechanism of motion of minority carriers in the base, transistors are grouped into drift and diffusion types. A diffusion transistor features a uniform distribution of impurities in the base (Fig. 7.1) and consequently the absence of an electric field therein. An example of the first type can be germanium alloy (fused) *p-n-p* transistors. They are prepared by fusing indium into two opposite

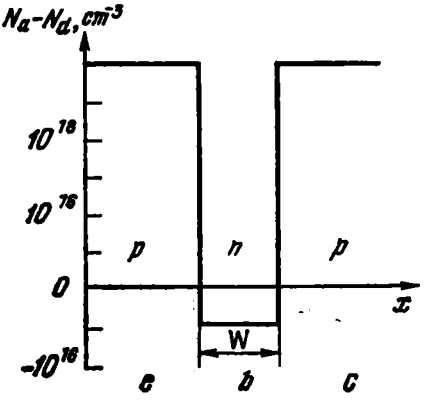


Fig. 7.1. Impurity distribution in an alloy transistor

sides of a germanium crystal (Fig. 7.2), the impurities being evenly distributed throughout its bulk.

Silicon transistors made by the method of sequential double diffusion of boron and phosphorus can serve as an example of drift transistors. Boron diffused into the original n-type silicon forms a p-region of a base and a collector p-n junction disposed at a depth x_c. Phosphorus is subsequently diffused into the crystal to a smaller depth x_e to form an emitter p-n junction. The temperature and duration of the two diffusion processes are chosen such as to obtain a predetermined base width $W = x_c - x_e$.

Figure 7.3 illustrates changes in the distribution of impurities in the semiconductor following each stage of diffusion. As a result, two p-n junctions are formed and in accordance with the law of diffused-impurity distribution a charge-carrier density gradient appears in the base. The majority carriers will naturally move in the direction which suits them most from the viewpoint of their energy levels, i.e., from the region of a higher impurity concentration to the lower concentration region. The carrier densities will continue to level off until they become offset by the electric field of the unbalanced electric charge induced due to the drift of majority carriers. The electric field so developed will impede the transport of majority carriers and, at the same time, will accelerate the transfer of the minority

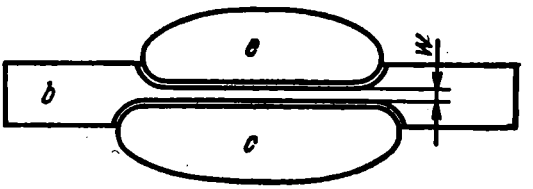


Fig. 7.2. Alloy p-n-p transistor structure

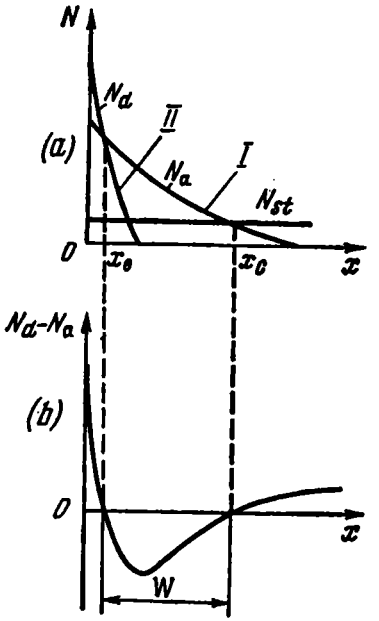


Fig. 7.3. Impurity distribution in a drift transistor prepared by the method of double diffusion

(*a*) *I*—diffused impurity distribution in forming the base region (first diffusion); *II*—diffused impurity distribution in forming the emitter region (second diffusion); (*b*) resultant impurity distribution

carriers injected by the emitter into the base region, the latter phenomenon being of particular importance.

The electric field adds to the drift velocity of minority carriers and thus lowers their transit time, improving thereby the frequency response of a transistor.

Since transistors are intended to perform various functions, they are distinguished by the purpose they serve.

Transistors are classed into amplifiers intended for amplifying electric signals, oscillators for generating oscillations, and switching transistors (or switches) forming a separate group.

Depending on their applications, transistors must exhibit a requisite set of electric parameters to be suitable for use in certain circuits.

Specifications for modern transistors stipulate more than twenty electric and other working parameters, the list of parameters as well as standards and demands on them being dictated by the purpose a particular device must serve and the potentialities of the manufacturer.

Transistors may be classified under various types by their main electric and other working parameters. Thus, as for their dissipated power, transistors are classed into low-power (usually up to 0.3 W), medium-power (0.3 to 3 W) and high-power (over 3 W) devices. Transistors are also divided into low-frequency (up to 10 MHz), medium-frequency (10 to 100 MHz), and high-frequency (over 100 MHz) devices by the value of the alpha cutoff frequency (in the common-base connection).

According to the USSR State Standard (ГОСТ) 10862-64, transistors are designated as follows. The first symbol (letter or digit) identifies the starting material a transistor is made of: 1 or Г for germanium, 2 or K for silicon, 3 or A for gallium arsenide. The second letter symbol stands for the class of a device: T (transistor). The third notation is a three-digit number, the first digit denoting the group the transistor belongs to by the value of frequency and dissipated power, and the last two digits designating the serial number of the series developed, namely:

low-power, low-frequency transistor 101 to 129
low-power, medium-frequency transistor 201 to 299
low-power, high-frequency transistor 301 to 399
medium-power, low-frequency transistor 401 to 499
medium-power, medium-frequency transistor 501 to 599
medium-power, high-frequency transistor 601 to 699
high-power, low-frequency transistor 701 to 799
high-power, medium-frequency transistor 801 to 899
high-power, high-frequency transistor 901 to 999

When the transistor series consists of a group of devices the major parameters of which are very similar though there are some deviations in certain parameters, these modified devices are differentiated among themselves by a letter symbol at the end of the notation. For example, 1T308A means a germanium transistor of low power and high frequency, version A; KT312B — low-power, high-frequency silicon transistor, version B.

Prior to 1964 transistors were denoted by letter П and a three-digit number in accordance with the following division:

low-power, low-frequency germanium transistors 1 to 99
low-power, low-frequency silicon transistors 101 to 199
high-power, low-frequency germanium transistors 201 to 299
high-power, low-frequency silicon transistors 301 to 399
low-power, high-frequency germanium transistors 401 to 499
low-power, high-frequency silicon transistors 501 to 599
high-power, high-frequency germanium transistors 601 to 699
high-power, high-frequency silicon transistors 701 to 799

The division values determining the grouping of transistors were: according to power — 0.25 W, and frequency — 5 MHz.

7.2. ALLOY TRANSISTORS

We now examine diffusion transistors differing in design and methods of preparation. It should be noted that the relations derived in Chapter Five are fully applicable for diffusion transistors.

The alloy transistor is a device whose *p-n* junction is formed by fusing an impurity into a semiconductor material. Prior to fusion, germanium or silicon wafers are carefully polished and etched to

insure the preset width and very accurately assorted as to their thickness. Alloying is usually made either in fixtures (jigs) or without fixtures on a conveyor belt.

In the fixture method, before being put into a furnace, a semiconductor wafer and doping pellets are fastened in an alloying jig, often made of graphite. The jig provides for the alignment and contact between the parts and also prevents the molten pellets from spreading about. Alloying may be done on the two opposite faces at the same time (emitter and collector junctions alloyed in simultaneously) or in succession: first, by forming the collector and then the emitter junction. The same jig is most commonly used to bond the crystal to the crystal holder, forming thereby contact with the base region.

In the alloy process carried out without jigs, semiconductor wafers are placed on a conveyor belt with holes through which doping pellets (electrode material) are inserted and pressed against the wafers. The pellets are then kept in position by cohesive force. This technique is suitable for the production of transistors whose pellets are so small that it is impossible to load them into jigs by usual methods.

Two alloying elements, namely indium and aluminium, are alloyed in their pure state: the former, when preparing germanium *p-n-p* transistors and the latter, silicon *p-n-p* transistors. In most cases the doping pellet (electrode alloy) is an electrically neutral material (as lead, gold, silver) with an addition of a doping element (as antimony, phosphorus, arsenic, gallium, aluminium) or an alloy consisting of many elements each intended to improve a particular property of a transistor. Thus, the use of Au-Pb-Sb or Sn-P alloys in the production of silicon *n-p-n* transistors yields good results.

Once prepared for alloying, the wafers with doping pellets are placed on a moving conveyor belt in an oven with a rigidly controlled range of specified temperatures and are subjected to heat treatment in conformity with the prescribed temperature-time cycle.

The main purpose of the alloying process is to obtain the required depth of impurity penetration securing thereby a predetermined distance between the collector and the emitter junctions, i.e., the base width. Of great importance here is a close control over the temperature and the time of alloying selected in each particular case for wafers of given thickness.

Wafer assemblies with fused junctions are checked by the shape of their current-voltage curves and the value of the current transfer ratio. The selected good structures are then fitted with leads and their surfaces are cleared of contaminants. The surface of alloy transistors is highly sensitive to the effect of an environmental medium, especially moisture, and for this reason they should be enclosed in metal-glass cases (envelopes) to provide for the dry and controlled atmosphere of the transistor. The transistor stem assem-

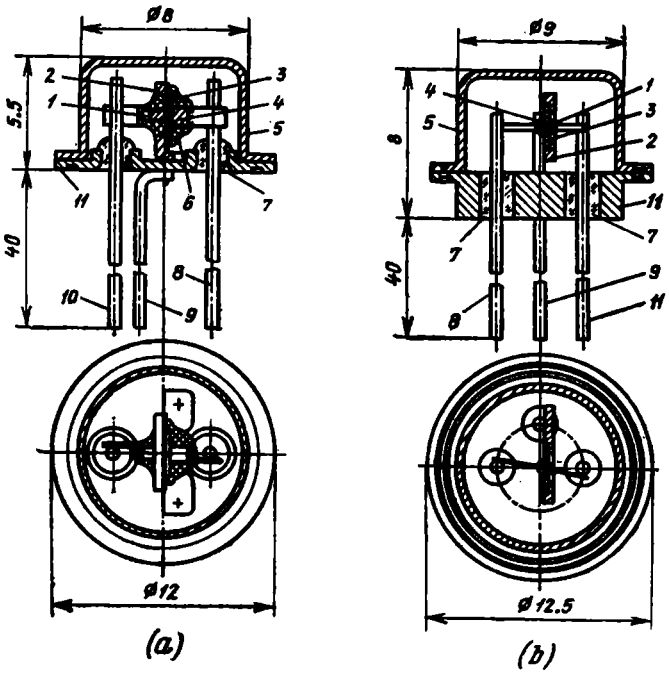


Fig. 7.4. Designs of low-power alloy transistors

(*a*) enclosure sealed by electric welding; (*b*) enclosure sealed by cold welding; *1*—emitter pellet; *2*—crystal holder; *3*—crystal; *4*—collector pellet; *5*—can; *6*—protective lacquer; *7*—insulator; *8*—collector lead; *9*—base lead; *10*—emitter lead; *11*—stem flange

bly with the wafer structure mounted thereon and the can (shell) are joined together by electric or cold welding. In the latter technique, the can and the stem flange, both made of ductile materials (copper and kovar), are joined under high pressure in specially shaped dies. The two parts are held together due to intermolecular cohesion developed under pressure forces. Cold welding enables one to achieve highly reliable encapsulation and to eliminate the effect of temperature on the crystal during welding, though it is not devoid of disadvantages which are the following: stringent requirements for cleanliness of the parts being welded and excessive expenditure of such expensive materials as oxygen-free copper and kovar, which greatly add to the cost of the device design.

The limitations of electric welding are thermal shock and liberation of harmful gases and particles during the passage of current and heating, its merits being lower cost and efficiency. The most typical encapsulation structures of low-power alloy transistors are illustrated in Fig. 7.4.

It is difficult to attain a sufficiently high percentage yield of serviceable transistors with a base width of less than 10-15 μm, and for this reason the cutoff frequency of alloy transistors is not higher than 20 MHz. The parameters of most popular low-power germanium and silicon alloy transistors turned out in the USSR are listed in Table 7.1.

Table 7.1

Parameters of Germanium and Silicon Alloy Transistors

Parameters	Germanium				Silicon	
	П9-П11А	МП13-МП15	П20-П21А	П25-П26В	П101-П103	П104-П106
	n-p-n	*p-n-p*	*p-n-p*	*p-n-p*	*n-p-n*	*p-n-p*
Current-gain cutoff frequency, MHz	1-2	0.5-2	1	0.2-0.5	0.2-1	0.1-0.5
Collector reverse current, μA	30	15	50	150	1	1
Emitter reverse current, μA	30	30	50	150	1	—
Collector capacitance, pF	60	50	—	50-70	—	80
Base resistance at high frequency, ohm	150	150	—	150	—	—
Maximum permissible collector voltage, V	15-30	15-30	15-70	60-100	10	15-60
Maximum permissible collector current, mA	20	20	300*	400*	20	10
Maximum permissible power, mW	150	150	150	200	150	150

* Of the current pulse.

High-power alloy transistors are distinguished from low-power ones primarily by the increased area and perimeter of *p-n* junctions. This increase is necessary for passing heavy currents and improving heat removal from the collector. It is common to mount the device structure directly on the flange of the can (Fig. 7.5).

Transistors of high power are provided with additional means for heat withdrawal from the cases. When air-cooled, the devices are fitted with special heat exchangers (metallic radiators) having a large surface of heat emission (Fig. 7.6). Natural convection of air

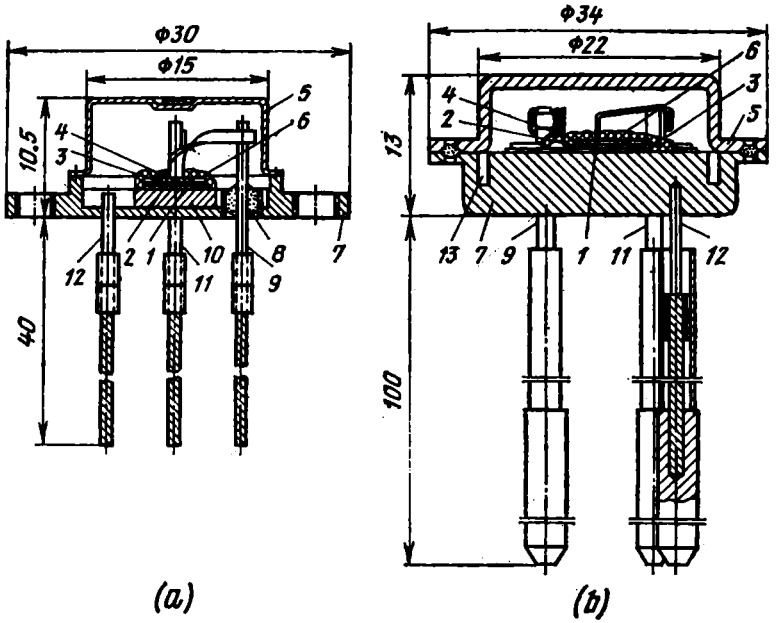


Fig. 7.5. Designs of high-power alloy transistors

(*a*) encapsulation by electric welding; (*b*) encapsulation by cold welding; *1*—collector; *2*—crystal; *3*—ring-shaped contact of the base; *4*—emitter; *5*—can; *6*—protective lacquer; *7*—stem flange; *8*—insulator; *9*—emitter lead; *10*—copper insert; *11*—base lead; *12*—collector lead; *13*—groove making up for deformation of the flange during encapsulation

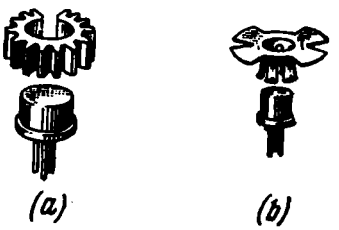


Fig. 7.6. Types of heat radiators for transistors

(*a*) star type; (*b*) wing type

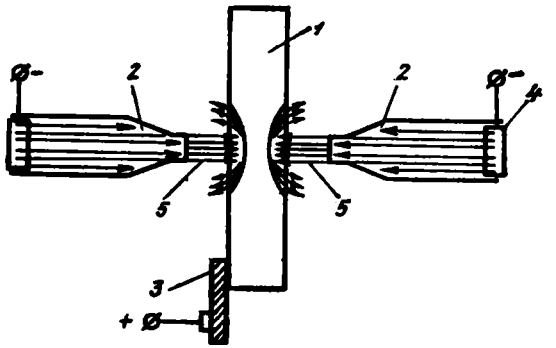


Fig. 7.7. Scheme of jet-type electrochemical etching

1—crystal; *2*—nozzle; *3*—crystal holder; *4*—contact; *5*—electrolyte jet

is sufficient to cool the transistors with a power of about 10 watts, while higher-power transistors require forced air cooling.

In common use is also a contact method of cooling by which the transistor case is fastened directly onto a metallic plate of high thermal conductivity. To avoid electric contact between the case and the plate and to fill in the poorly heat-conductive air gap between the contact surfaces, the latter are greased with a special insulating paste or are separated by a thin interlayer of insulation. Forced liquid cooling is sometimes used, which is more effective than air cooling but is complicated and therefore less popular.

In exceptional cases transistors are cooled by thermoelectric coolers. This method is rather complex and costly. Its advantage over other methods consists in that the temperature of the transistor case may be held lower than that of the environment. The microalloy transistor is a variety of the alloy transistor. To obtain a sufficiently thin base for the microalloy transistor and consequently raise its cutoff frequency, the working portion of the crystal is shaped by electrochemical jet etching (Fig. 7.7). The process is easily controlled by adjusting the nozzle diameter, speed of electrolyte outflow and current passed through the system. This makes it possible to obtain a very thin bridge between the holes etched off in the starting material. On depositing by the electrochemical method a thin layer of indium on and fusing it in the bottom of the holes one can obtain a base width in the order of 5 μm, and thus raise the current-gain cutoff frequency to several tens of megahertz.

Owing to their comparatively low cutoff frequency, small power of dissipation (not more than 10-15 mW) and elaborate fabrication, the microalloy transistors cannot compete with high-frequency tran-

sistors produced by diffusion methods. At present the production and use of microalloy transistors are rather limited.

One should note that the alloy transistors are inferior to the diffused-junction transistors in their characteristics and are gradually supplanted by the latter. However, alloy transistors are as yet most popular owing to ease of fabrication and low cost.

7.3. DRIFT TRANSISTORS

Drift transistors feature an electric field in the base, caused by the density gradient of impurities. Nonuniform distribution of impurities in the bases of these transistors is attained by the diffusion technique employed for their production.

Since the diffusion process is much slower and better controllable than the alloying process, it ensures the production of transistors with an extremely thin base and thus makes possible an appreciable increase in the cutoff frequency. If the cutoff frequency of alloy transistors reaches several, at most one or two tens of megahertz, the cutoff frequency of diffused transistors runs as high as hundreds and even thousands of megahertz.

In the production of high-frequency germanium transistors with a current-gain cutoff frequency of several hundreds of megahertz, wide use is made of a method combining diffusion with alloying. Now consider the method of preparation of germanium diffused-alloy transistors in more detail.

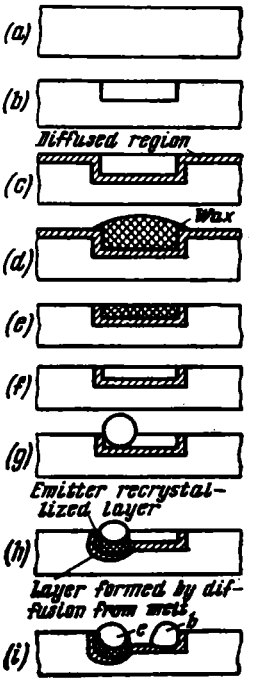


Fig. 7.8. Steps of fabrication of П401—П403 low-power diffused-alloy transistors

(*a*) slice lapping; (*b*) grooving and etching; (*c*) antimony diffusion; (*d*) groove protection by wax; (*e*) slice lapping; (*f*) wax removal; (*g*) insertion of emitter pellet; (*h*) fusion and diffusion; (*i*) insertion and fusion of base pellet

In a slice of *p*-germanium with resistivity $\rho = 0.5\text{-}1$ ohm·cm grooves are made by ultrasonic vibration dimensioned to 0.4 mm in length, 0.1 to 0.12 mm in width, 0.1 to 0.15 mm in depth and spaced 1 to 2 mm apart (Fig. 7.8). An *n*-germanium layer 0.02 mm thick is then formed by diffusing antimony. This layer is afterwards so lapped or etched that the *n*-germanium layer is left intact only in the grooves. The grooves are protected against etching by filling them with wax or special varnish.

Next, two pellets of an In-Ga-Sb-Au alloy (in the proportion 88:1.5; 3.5:7%) and a PbSb alloy (95:5%) are alloyed in each groove in succession. After the pellets have been fused in the hydrogen

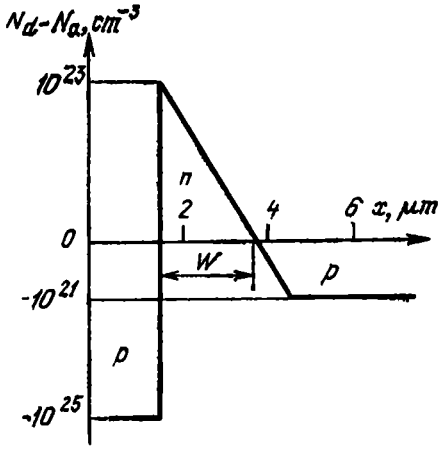


Fig. 7.9. Impurity distribution in a diffused-alloy transistor

atmosphere, the temperature is slightly lowered and the structure is held for some time at that temperature. Lowering the temperature and keeping it for a certain time result in the formation of a recrystallized layer from which antimony, gallium and indium are diffused underneath the first drop (pellet) and only antimony underneath the second drop (pellet). As a result, a transistor structure is formed because antimony has a higher diffusion constant and lower ultimate solubility than indium and gallium. The first pellet serves as an emitter electrode and the second as a base one which is connected to the thin active base region lying under the emitter pellet with a bridging layer formed earlier by the deposition and diffusion of antimony. The resultant distribution of impurities in the diffused-alloy transistor is shown in Fig. 7.9.

The sequential diffusion methods are used in the production of planar and mesa transistors. They are fabricated by localized selective diffusion with oxide masking. A distinctive characteristic of planar transistors is the in-line disposition of the boundaries of both *p-n* junctions (emitter-base and collector-base junctions), as is shown in Fig. 7.10. The main principles of the planar technique employed in

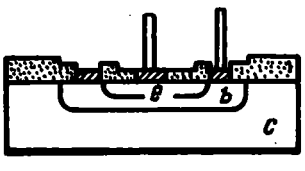


Fig. 7.10. Planar transistor structure

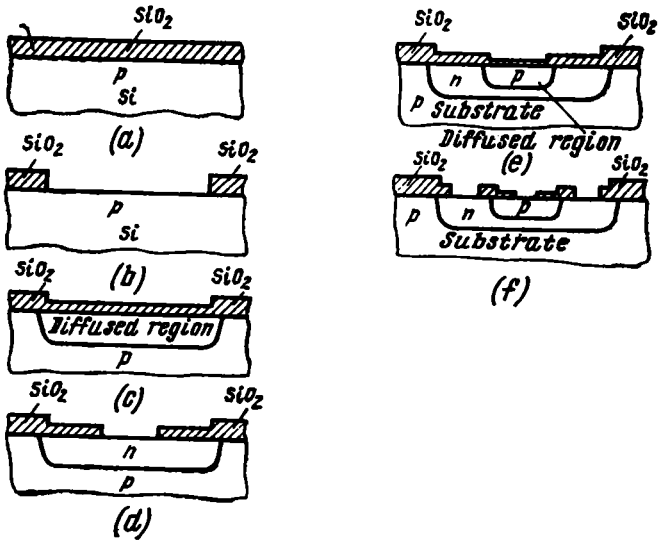


Fig. 7.11. Flowchart of fabrication of a planar transistor

(*a*) oxide-coated slice; (*b*) preparation of the slice to form the base by diffusion; (*c*) structure with a collector *p*-layer; (*d*) slice prepared for diffusion of an emitter impurity; (*e*) structure with an emitter and a collector junction; (*f*) structure prepared for deposition of ohmic contacts onto the emitter and the base

the production of planar transistors will be described below. The steps of fabrication of such transistors are shown in Fig. 7.11.

An initial silicon oxide film is grown on the surface of a silicon wafer, for example, by thermal oxidation (Fig. 7.11*a*). Then a thin layer of special photosensitive material, the so-called photoresist, stable to an etchant, such as hydrofluoric acid, is deposited on the silicon oxide film. Further, the structure is exposed to light through a special mask and is subsequently developed and heat treated to obtain holes (windows) in the photolayer, their number and configuration corresponding to the pattern or design on the photomask. Etching the silicon oxide by hydrofluoric acid results in the reproduction of the pattern on the silicon oxide, while silicon remains intact as it does not dissolve in hydrofluoric acid (Fig. 7.11*b*). The whole combination of these processes is identified as photolithography, photomasking or photoengraving on oxide.

During the diffusion process silicon is alloyed only at the spots where the oxide has been etched away owing to a great difference in magnitude between the diffusion constants of silicon and silicon oxide (Fig. 7.11*c*). Following the first diffusion step carried out to form a base region, the silicon is oxidized again and the obtained

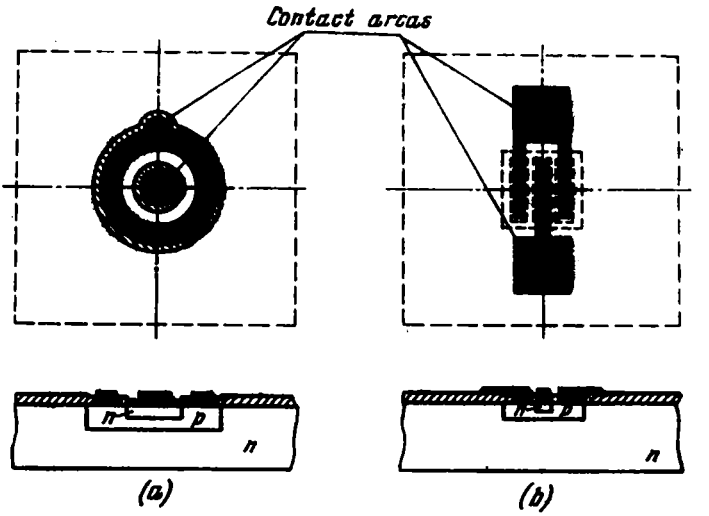


Fig. 7.12. Configuration of planar transistors
(a) ring-shaped type; (b) strip type

oxide is then photomasked (Fig. 7.11*d*). Then another step of diffusion of an alloying impurity follows to form an emitter region. After the subsequent oxidation all the *p-n* junction boundaries that reach the surface become coated with an oxide layer (Fig. 7.11*e*). This ensures more stable properties of the surface and small reverse currents. The final step includes the provision of appropriate windows by photomasking (Fig. 7.11*f*), into which a metal is deposited to form reliable ohmic contacts for the subsequent attachment of leads (Fig. 7.10).

The geometry of transistor regions and that of contact areas are conditioned by photomasks and are usually of a circular or a strip type (Fig. 7.12). It is evident from Fig. 7.12 that if the emitter region has a ring configuration, it cannot be in principle smaller than the contact area. The strip configuration enables one to bring out the contact areas over and above the masking oxide in a convenient manner and thus to cut down the size of the active regions of a transistor and consequently to decrease the collector and emitter capacitancies.

A characteristic feature of the mesa transistor is the fact that its active regions form a projection on the semiconductor crystal surface (Fig. 7.13). This projection is formed when portions of the wafer are etched away in order to appreciably reduce the passive region of the base. Since the geometry of the collector-base junction depends on

the size of the protective mask used during etching, it is not mandatory to carry out localized diffusion of a base impurity, such as gallium, for which the silicon oxide is not a masking substance. The flowchart of fabrication of a mesa transistor is given in Fig. 7.13. A semiconductor crystal, silicon, for example, is alloyed in by boron or gallium. The surface is then masked by silicon oxide in which "windows" are subsequently made by the photoengraving method to diffuse impurities through the windows and thus form an emitter junction. On completion of the diffusion process, the wafer is coated with a masking composition, for example, wax or photoresist, and then the so-called mesa projection of the desired configuration is made by etching.

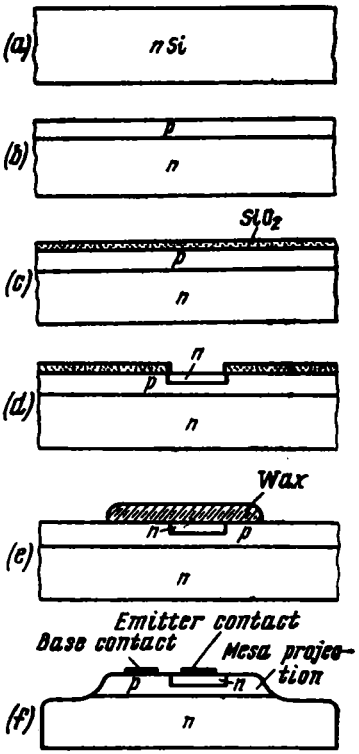


Fig. 7.13. Steps of fabrication of a mesa transistor

(*a*) starting silicon slice; (*b*) base impurity diffused into silicon; (*c*) slice coated with oxide; (*d*) emitter impurity diffused into silicon; (*e*) structure coated with wax and ready for etching; (*f*) ready-made transistor structure

Planar and mesa transistors resemble alloy transistors in appearance as they are encapsulated in unitized enclosures, the only difference being that both small-power and high-power planar transistors are usually assembled with their collectors directly attached to the cans. This is dictated by the peculiarities of the structure formed in the starting material containing the collector. When it is necessary to isolate the collector from the can, an insulating and, if possible, a well heat-conducting material, for example, beryllium oxide-based ceramics, is placed between them. Rather extensive use has been made recently of the so-called caseless silicon planar transistors. The transistor structures are safeguarded against exposure to the atmosphere by a thin protective layer made, for example, of glass or special resin-based compound, or, sometimes, by a multilayer coating. Such a coating enables the transistor structure to withstand rather long-term exposure to the environment and protects it from damage in transit and storage before its mounting in a circuit unit. The advantage of these transistors is their extremely small size (Fig. 7.14) and low dissipated power.

One of the essential parameters of high-frequency fast-response

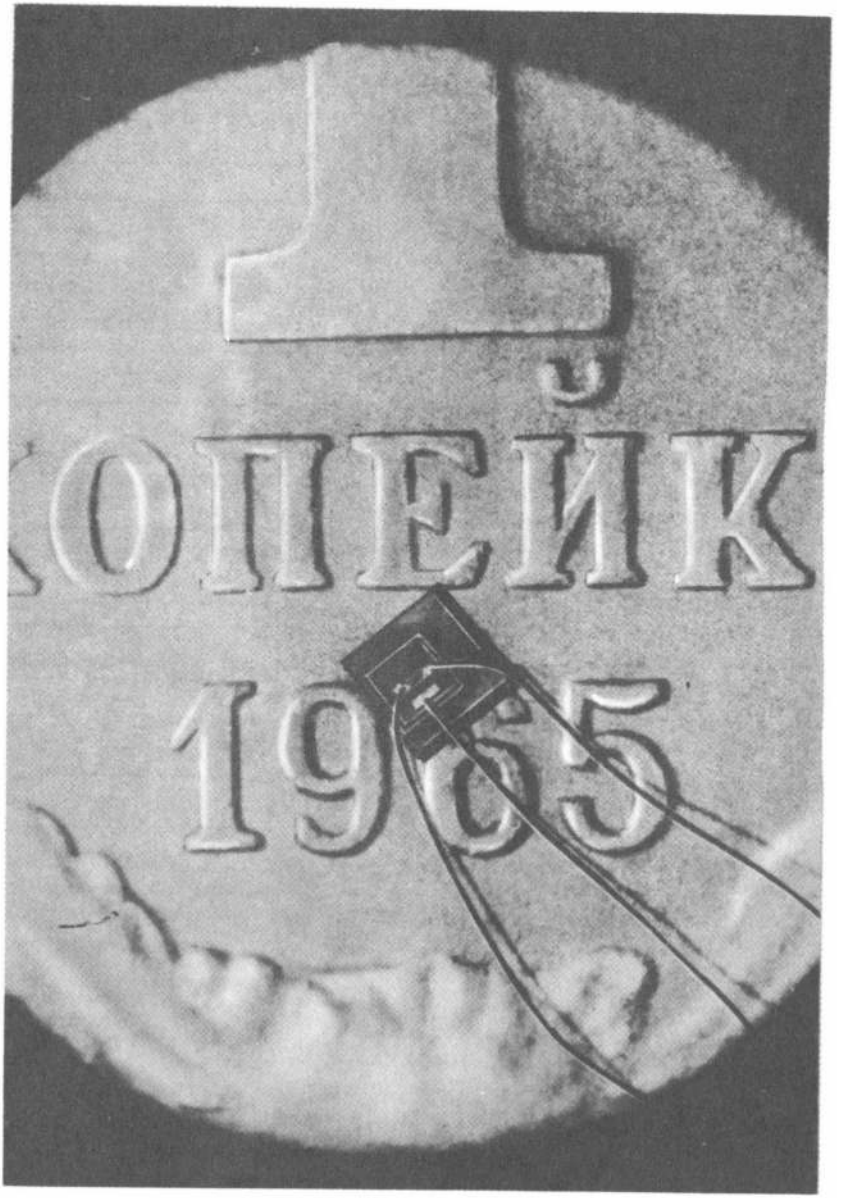


Fig. 7.14. **Caseless planar transistor**

transistors is the residual resistance of the collector in the saturated mode of operation. The lower this resistance, the better the performance of transistors in pulse circuits. In planar transistors, as opposed to alloy transistors, the residual resistance is a function of the size and conductivity of the bulk of the collector region.

Indeed, all the active regions of the transistor are disposed in a layer lying 5 to 10 μm under the surface. But it is impossible to handle crystals less than 100 μm thick due to their low mechanical strength. The resistance of the silicon crystal over 5 to 10 μm thick is considered parasitic. The choice of the resistivity of the starting material depends on the magnitude of the required breakdown voltage and allowable collector-junction capacitance. This resistivity usually reaches 0.2 to 1 ohm·cm.

To reduce the parasitic resistance of the crystal bulk, recourse is made to special alloying by the methods of out (outward) diffusion and epitaxial growing for obtaining a layer structure. The first technique consists in preliminary diffusion of an impurity that does not change the type of majority carriers deep into the wafer, providing at the same time a high surface density. In the fabrication of *n-p-n* transistors phosphor is used as a diffusate, which is readily soluble in silicon and has a high diffusion constant.

Following the diffusion, one side of the silicon wafer is lapped plane-parallel to the other to remove completely the n^+-layer and reduce to a minimum the original *n*-layer so that it can accommodate only the base and emitter active regions to be formed by diffusion. Excessive lapping may result in the displacement of the collector-base region into the n^+-region and thus cause a reduction of the breakdown voltage.

This imposes stringent demands on the accuracy of lapping (required size and plane-parallelism should be accurate to ±2 μm).

The out diffusion method permits lowering the saturation resistance or reducing the residual voltage across the collector by 75-87 per cent.

Let us assess the saturation resistance which depends on the collector region thickness. Fig. 7.15 shows the distribution of impurities in a transistor structure formed without outward diffusion (*a*) and with the use of outward diffusion (*b*). The resistance of the original crystal layer made without outward diffusion is

$$r = \frac{\rho_m (W_c' + W_c'')}{S} \tag{7.1}$$

where ρ_m is the resistivity of the material; $W_c' + W_c''$ is the collector region thickness; and S is the collector region area. Thus, when $\rho_m = 1$ ohm·cm, $W_c' = 10$ μm, $W_c'' = 120$ μm, and $S = 10^{-4}$ cm² (100 × 100 μm²)

$$r = \frac{1 \times 10^{-3} + 1 \times 1.2 \times 10^{-2}}{10^{-4}} = 130 \text{ ohms}$$

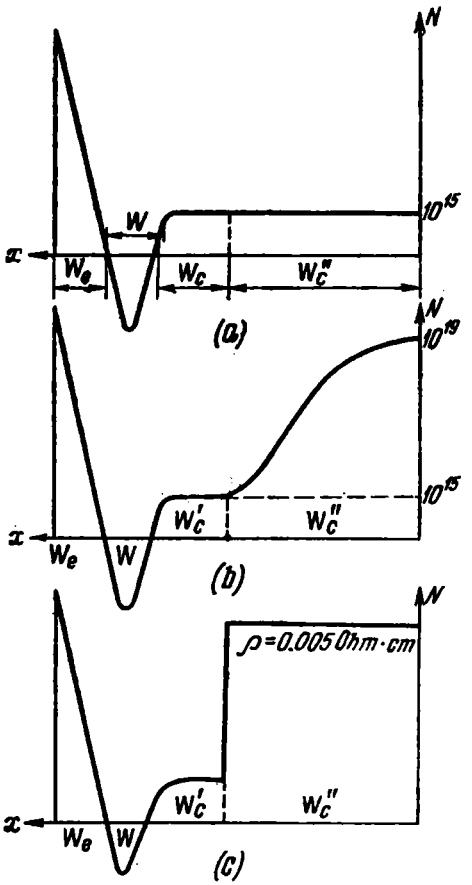


Fig. 7.15. Impurity distribution in transistor structures made without outward diffusion (*a*), with outward diffusion (*b*), and with an epitaxial collector (*c*)

The resistance of the collector formed with the aid of out diffusion is

$$r = r_1 + r_2$$

where

$$r_1 = \frac{\rho_m \cdot W_c'}{S} = \frac{1 \times 10^{-3}}{10^{-4}} = 10 \text{ ohms};$$

$$r_2 = \int_0^W \frac{dx}{q\mu(x) N(x) S} \qquad (7.2)$$

This integral may be approximated to the expression

$$r_2 = \frac{\rho_m}{2 \ln \frac{N_0}{N_m}} \frac{W''}{S} \qquad (7.3)$$

if we assume that $\mu(x) \approx \mu_m$, i.e., the mean value of mobility.

Hence, with the concentration of impurities in the original material, $N_m = 10^{15}$ cm^{-3}, the maximum impurity solubility $N_0 = 10^{19}$ cm^{-3}, and $W_c'' = 120$ μm

$$r_2 = \frac{1 \times 1.2 \times 10^{-2}}{2 \times 2.3 \log \left(\frac{10^{19}}{10^{15}}\right) \times 10^{-4}}$$

$$\approx 6.5 \text{ ohms}$$

The total saturation resistance of the collector region $r_{cs} = 10 + 6.5 = 16.5$ ohms, which is one eighth of that of the collector region made without outward diffusion. If we take into account the resistances of the junctions and the base region the difference will be less impressive.

We can benefit still more from using silicon wafers with an epitaxially grown layer. It will be recalled that the epitaxial growing is the process of growing on the substrate a crystal layer with a crystalline structure which continues the structure of the substrate.

Table 7.2

Parameters of Low-Power High-Frequency Germanium and Silicon Transistors

Parameters	Germanium				Silicon			
	encased						caseless	
	diffused-alloy		mesa	mesa	planar epitaxial			
	П401-П403А, *p-n-p*	П416-П416Б, *p-n-p*	1Т308-1Т308Г, *p-n-p*	П504-П505, *n-p-n*	2Т306А—Г, *n-p-n*	2Т316А—Г, *n-p-n*	2Т307А—Г, *n-p-n*	2Т324А—Д, *n-p-n*
Cutoff frequency, MHz	30-120	40-80	150-200	50-150	300-500	600-800	250	600-800
Collector reverse current, μA	5-10	3	0.4	2	0.5	0.5	0.5	0.5
Collector capacitance, pF	10-15	8	—	7	5	3	6	2.5
Emitter capacitance, pF	—	—	—	—	4.5	2.5	—	—
Feedback time constant, ps	500-3500	500	100-500	—	—	150	—	—
Storage time, ns	—	1000	—	—	30	10	30	15
Maximum collector-emitter voltage, V	10	12	12	20-30	10	10	10	10
Maximum emitter-base voltage, V	—	3	2-3	2	4	4	4	4
Maximum collector current, mA	10	15	50	10	30	30	20	20
Maximum power, mW	50	100	150	150	150	150	15	15

To reduce the saturation resistance, a 6 to 10-μm thick epitaxial layer of a required high resistivity is deposited or grown on the low-resistivity substrate 100 to 200 μm in thickness and with $\rho_m \approx \approx 0.005$ ohm·cm, a planar or mesa transistor structure being subsequently formed in the epitaxial layer. The series resistance of a collector bulk and consequently the voltage drop across the collector region of a planar or mesa transistor formed on the wafer with an epitaxial layer are usually one third to one tenth of these quantities in the transistor fabricated by the common method (see Fig. 7.15*c*). Indeed, if a 10-μm epitaxial layer with $\rho = 1$ ohm·cm is grown on a 120-μm substrate with $\rho_m = 0.005$ ohm·cm, then

$$r_{cs} = r_1 + r_2 = 10 + \frac{5 \times 10^{-3} \times 1.2 \times 10^{-2}}{10^{-4}} = 10.6 \text{ ohms}$$

i.e., one twelfth of the saturation resistance in the common transistor and two thirds of that in the transistor formed by out diffusion. An added advantage of such transistors, called planar epitaxial transistors, is the improvement of device characteristics.

The designer used to compromise in selecting the original material of a required resistivity because, on the one hand, a higher value of ρ is conducive to a higher breakdown voltage and lower capacitance, on the other hand, a higher resistivity leads to considerable power losses. Now, by using the epitaxial technique it is possible to grow a layer with a significantly high resistivity and thus have a margin for variations in voltage and capacitance without fearing too sharp a rise in parasitic loss. At present most planar transistors are fabricated with the use of the epitaxial technique.

The speed of response and the operating frequencies of planar transistors are extremely high. The cutoff frequency of low-power planar epitaxial transistors turned out by Soviet industry runs as high as 2 GHz and that of high-power transistors, hundreds of megahertz with the dissipated power coming to several watts. The parameters of the most popular types of low-power high-frequency germanium and silicon drift transistors are given in Table 7.2.

7.4. CALCULATION OF THE PLANAR TRANSISTOR

Let us now exemplify an approximate calculation of planar transistor parameters. As an example we may take an *n-p-n* transistor structure shown in Fig. 7.12*b*. Dimensions and other initial data required for the calculation are given below:

Surface donor impurity density in the emitter	$N_{de} = 10^{20}$ cm^{-3}
Acceptor impurity density in the base near the emitter	$N_{abe} = 10^{17}$ cm^{-3}

Acceptor impurity density in the base near the collector	$N_{abc} = 10^{15}$ cm^{-3}
Acceptor impurity density in the emitter near the surface	$N_{ae} = 10^{20}$ cm^{-3}
Base width	$W = 2 \times 10^{-4}$ cm
Electron lifetime in the base	$\tau_{nb} = 5\mu s$
Hole lifetime in the emitter	$\tau_{pe} = 0.1\mu s$
Emitter junction area	20×100 μm^2 or 2×10^{-5} cm^2
Collector junction area	150×150 μm^2 or 2.25×10^{-4} cm^2

In many practical cases the impurity distribution after diffusion is considered exponential for simplifying calculations:

$$N(x) = N_0 e^{-ax} \tag{7.4}$$

The current gain in the common-base connection is equal to

$$h_{21b} = \alpha = \beta^* \gamma^*$$

where β^* and γ^* are calculated by formulas (5.17) and (5.20).

To determine γ^* it is necessary to know the diffusion constant for electrons in the base and that for holes in the emitter.

For the electrons in the base the mean diffusion constant is determined from the Einstein relation (1.8):

$$D_{nm} = \frac{kT}{q} \mu_m$$

In its turn,

$$\mu_m = \frac{W}{t_{nb} E} \tag{7.5}$$

where t_{nb} is the time of transition of electrons across the base layer. According to the definition,

$$t_{nb} = \int_0^W \frac{dx}{v_p(x)} = \int_0^W \frac{dx}{\mu(x) E} \tag{7.6}$$

We assume that the field is constant and the mobility

$$\mu(x) = \mu_{nbe} + bx \tag{7.7}$$

where μ_{nbe} is the electron mobility in the base near the emitter. As is proved by experiment, the magnitude of b entering into (7.7) is little dependent on the coordinate x. Hence,

$$t_{nb} = \int_0^W \frac{dx}{(\mu_{nbe} + bx) E} = \frac{1}{bE} \ln \left(\frac{\mu_{nbe} + bW}{\mu_{nbe}} \right) \tag{7.8}$$

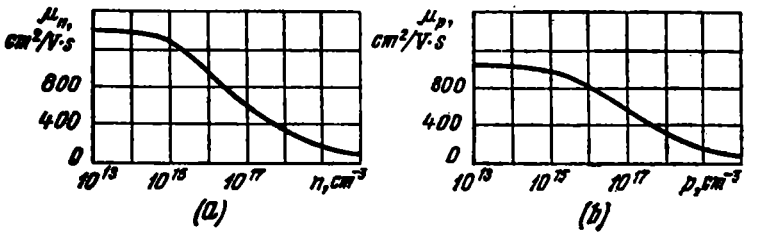


Fig. 7.16. Mobility of majority carriers in silicon as a function of carrier density at 300°K
(a) n-type silicon; (b) p-type silicon

Correspondingly,

$$\mu_m = \frac{bW}{\ln \frac{\mu_{nbe} + bW}{\mu_{nbe}}}$$

Proceeding from the dependence μ (N_a),

$$\mu_{nb} = 300\ (19 - \log N_a)$$

and assuming that N_a in the base varies from $N_{abe} = 10^{17}$ cm^{-3} to 2×10^{15} cm^{-3}, which corresponds to the base resistivity $\rho = 4.5$ ohm·cm, we obtain

$$b = \frac{300}{W} \log \frac{N_{abe}}{N_{det}} = \frac{300}{W} \log \frac{10^{17}}{2 \times 10^{15}}$$

whence $bW = 510$ cm/V·s.

Consequently,

$$D_{nbm} = \frac{kT}{q} \frac{bW}{\ln \frac{\mu_{nbe} - bW}{\mu_{nbe}}} = 0.026 \frac{510}{\ln \frac{600 + 510}{600}} = 25.5 \text{ cm}^2/\text{s} \qquad (7.9)$$

The diffusion constant D_{pe} for holes in the emitter when $N_d = 10^{20}$ cm^{-3} is easily determined by the value of $\mu_{pe} = 70$ cm^2/V·s from the graph of Fig. 7.16.

$$D_{pe} = \frac{kT}{q} \mu_{pe} = 0.026 \times 70 = 1.82 \text{ cm}^2/\text{s}$$

Hence, bearing in mind that

$$L_{pe} = \sqrt{D_{pe}\tau_{pe}} \quad \text{and} \quad L_{nbm} = \sqrt{D_{nbm}\tau_{bm}}$$

it is easy to find γ^*

$$\gamma^* = 1 - \frac{D_{pe}N_{abe}W}{D_{nm}N_{de}L_{pe}} = 1 - \frac{25.5 \times 10^{17} \times 2 \times 10^{-4}}{1.82 \times 10^{20} \times 4 \times 10^{-4}} = 1 - 0.0122 = 0.9878$$

and β^*

$$\beta^* = 1 - \frac{1}{2}\frac{W^2}{L_{nb}^2} = 1 - \frac{1}{2}\frac{(2\times 10^{-4})^2}{25.5\times 5\times 10^{-6}} = 0.99984$$

The current gain in the common-base circuit arrangement is equal to

$$h_{21b} = \alpha = \beta^*\gamma^* = 0.9998 \times 0.9878 = 0.9876$$

The current transfer ratio in the common-emitter connection is

$$h_{21e} = \beta = \frac{\alpha}{1-\alpha} = \frac{0.9876}{0.0124} = 78$$

The *transit time* can be found from (7.8) after the estimation of field strength E in the base, developed due to the nonuniform distri bution of impurities.

The approximate value of E may be given by

$$E = \frac{kT}{q}a \tag{7.10}$$

where a is the impurity density gradient, whence, according to (7.8) and (1.8),

$$t_{nb} = \frac{W}{aD_{nb}} \tag{7.11}$$

Let us determine the value of a from (7.4) by setting $x = W$:

$$a = \frac{1}{W}\ln\left(\frac{N_{abe}}{N_{abc}}\right) \tag{7.12}$$

As a result,

$$t_{nb} = \frac{W^2}{D_{nb}\ln\left(\frac{N_{abe}}{N_{abc}}\right)} \tag{7.13}$$

Comparing this expression with (5.70), we find that the transit time for holes in the drift transistor is $\ln\frac{N_{abe}}{N_{abc}}$ times less than that in the diffusion transistor. For our case,

$$t_{nb} = \frac{(2\times 10^{-4})^2}{25.5\times 10^{-3}\ln\frac{10^{17}}{2\times 10^{15}}} \cong 4\times 10^{-10}\ \text{s}$$

The *current-gain cutoff frequency* in the common-base arrangement is determined from the expression

$$f_{h21b} = f_\alpha = \frac{D_{nbm}}{\pi W^2}\left[1 + 0.85\left(\frac{qEW}{2kT}\right)^{3/2}\right] \tag{7.14}$$

From (7.10) and (7.12) we find

$$E=0.026\frac{1}{2\times10^{-4}}\ln\left(\frac{10^{17}}{2\times10^{15}}\right)=508\ \text{V/cm}$$

and

$$f_{h21b}=f_\alpha=\frac{25.5}{\pi(2\times10^{-4})^2}\left[1+0.85\left(\frac{39\times508\times2\times10^{-4}}{2}\right)^{3/2}\right]=686\ \text{MHz}$$

The *collector junction charge capacitance* is determined as the capacitance of a parallel-plate capacitor

$$C_{ch}=\frac{S\varepsilon\varepsilon_0}{\delta}=S\left[\frac{q(\varepsilon\varepsilon_0)^2N'}{12(\Delta\varphi_d-V)}\right]^{1/3} \tag{7.15}$$

V_c across the collector is commonly much smaller than $\Delta\varphi_d$. Therefore, considering that in our case $N'=aN_{dc}$, where

$$a=\frac{1}{W}\ln\frac{N_{abe}}{N_{abc}}=\frac{1}{2\times10^{-4}}\ln\left(\frac{10^{17}}{2\times10^{15}}\right)=1.95\times10^4\ \text{cm}^{-1}$$

and

$$N_{dc}=2\times10^{15}\ \text{cm}^{-3}$$

we have

$$C_{ch}=0.015^2\left[\frac{1.6\times10^{-19}\times8.86\times10^{-14}\times12\times1.95\times10^4\times2\times10^{15}}{12\times5}\right]^{1/3}=$$

$$=1.1\times10^{-12}\ \text{F}=1.1\ \text{pF}\ \text{when}\ V_c=5\ \text{V}$$

For the case when the surface impurity density is appreciably high and the emitter junction is disposed close to the surface, the *emitter junction capacitance* is calculated in the same way as for the abrupt junction according to (2.53), where

$$\Delta\varphi_d=\Delta\varphi_{deb}=\frac{kT}{q}\ln\frac{N_{de}N_{abe}}{n_i^2} \tag{7.16}$$

$$V=V_e=\frac{kT}{q}\ln\left(\frac{I_eWN_{abe}}{2qS_eD_{nbm}n_i^2}\right) \tag{7.17}$$

$$\Delta\varphi_{deb}-V_e=\frac{kT}{q}\ln\frac{2qS_eD_{nbm}N_{abe}}{I_eW} \tag{7.18}$$

when $I_e=5$ mA,

$$\Delta\varphi_{deb}-V_e=0.026\ln\frac{2\times1.6\times10^{-19}\times2\times10^{-3}\times10^{-2}\times25.5\times10^{20}}{5\times10^{-3}\times2\times10^{-4}}=$$

$$=0.252\ \text{V}$$

$$C_e=2\times10^{-3}\times10^{-2}\left[\frac{1.6\times10^{-19}\times12\times8.84\times10^{-14}\times10^{17}}{2\times0.252}\right]^{1/2}=$$

$$=3.68\times10^{-12}\ \text{F}=3.68\ \text{pF}$$

The reverse current and breakdown voltage in the collector and emitter junctions of drift and alloy transistors are calculated in a similar manner. Expressions for the calculation have been given above in Sec. 5.4.

Review Questions

1. Name the varieties of drift transistors and plot the graph of impurity distribution in these transistors.
2. What methods of encapsulation of alloy transistors do you know?
3. Why do planar transistors exhibit smaller reverse currents and increased stability?
4. Explain the process of forming the base region of a germanium diffused-alloy transistor.
5. What are the advantages of planar epitaxial transistors?
6. Determine the current-gain cutoff frequency f_{h21} for a silicon planar *n-p-n* transistor by setting $W = 1\mu m$ and using other initial data given in Sec. 7.4.
7. Which of the two types of silicon transistors, the *p-n-p* or *n-p-n* type, will have a higher cutoff frequency with one and the same width of the base, with other conditions being equal?

Chapter Eight

FIELD-EFFECT AND UNIJUNCTION TRANSISTORS. *p-n-p-n* SWITCHES

8.1. FIELD-EFFECT TRANSISTOR

The principle of operation of a channel-type transistor with a control electrode (gate) made in the form of a reverse-biased *p-n* junction was first advanced by Shockley in 1952 and realized by Pearson in 1953.

Let us examine a wafer made of semiconductor material with two *p-n* junctions and ohmic contacts attached to each of the three regions of different conductivity type disposed as is shown in Fig. 8.1. The electrodes located on the opposite ends of the medium region, of the *n*-type for the case being considered, are called source and drain in conformity with the direction of the flow of majority carriers. The two *p*-type regions flank the *n*-region that forms a current-conducting channel, and thus play the role of a gate.

Resistance R of the conducting channel is given by

$$R = \frac{l}{2q\mu_n nWd} \tag{8.1}$$

where W is the thickness of the channel; l is the channel length; and d is the channel width.

Now we connect the gate to the external voltage source so that both the *p-n* junctions become biased in the reverse direction (Fig. 8.2).

When the hole density in the *p*-regions considerably exceeds the electron density in the *n*-region, the subsequent expansion of the space charge layer in each *p-n* junction by the value δ will lead to the reduction in the thickness of the *n*-type bar by 2δ. This change in the thickness of the channel is determined by the magnitude of the voltage applied to the gate since according to formula (2.16) the value of $\delta \sim V^{1/2}$.

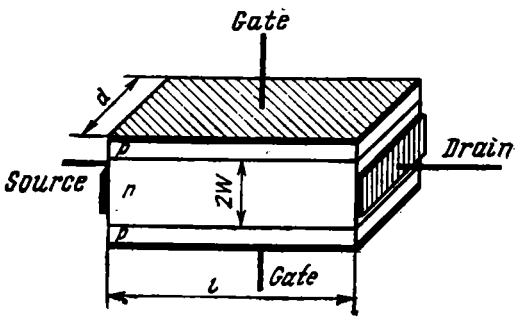


Fig. 8.1. Simplified field-effect transistor structure with a channel formed by *p-n* junctions

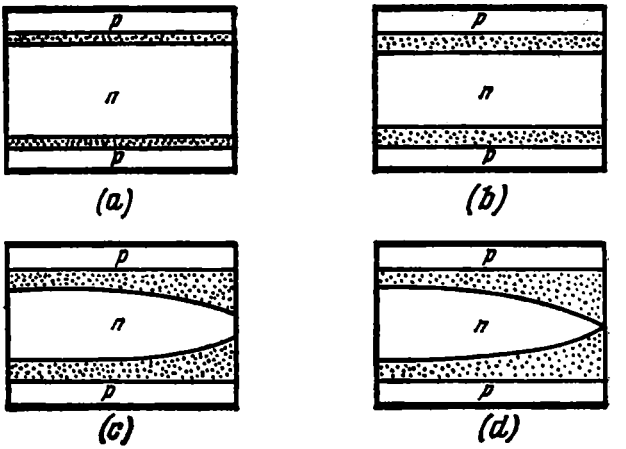


Fig. 8.2. Principle of operation of a field-effect transistor

(a) at gate voltage V_{g1} and no current through the channel; (b) at gate voltage $V_{g2} > V_{g1}$ and no current through the channel; (c) at gate voltage V_g and drain voltage V_d, where $V_g + V_d < V_0$; (d) $V_g + V_d \geqslant V_0$

It should be noted that at a certain value of the voltage applied to the gate the channel across its section can be entirely blocked by the expanded space charge layers of the *p-n* junctions. This voltage, called cutoff or pinch-off voltage, may be found from the condition $2\delta = 2W$ and by (5.59) equals

$$V_0 = \frac{W^2}{2\varepsilon\varepsilon_0\mu_n\rho_n}$$

The reverse bias applied to the gate tends to pinch off the channel, increase its resistance and decrease the working current. The voltage drop across the load resistance R_l series-connected with the channel begins to diminish correspondingly. Lowering the bias on the gate widens the channel and thus increases the current and the voltage drop across the R_l.

If we consider that the current flowing through the reverse biased *p-n* junction is very small compared to the channel current, then the possibility of amplification of the voltage and power of the signal applied to the gate becomes quite obvious.

Let us examine variations in the current-voltage characteristics of a channel-type transistor with changes of the voltage applied to its electrodes. In the absence of current through the channel and with a certain constant voltage across the gate, the potential across the channel throughout its length is invariable (Fig. 8.2*a*). With an increase of the voltage the thickness of the channel begins to dimi-

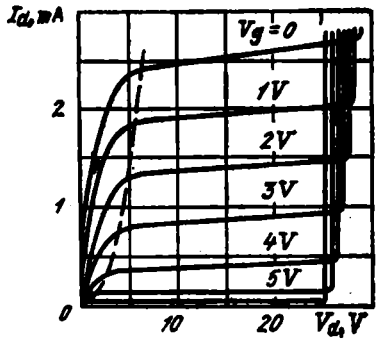


Fig. 8.3. Set of current-voltage curves of a field-effect transistor

nish uniformly along its length (Fig. 8.2*b*). The current passing through the channel causes a voltage drop which changes the shape of equipotential surfaces inside the channel. The potential across the channel has now become different at different points along its length and varies from zero at the source end to V_d at the drain end (Fig. 8.2*c*). The thickness of the channel has correspondingly tapered off in the direction from the source to the drain. This change determines the dependence of drain current I_d on drain voltage V_d, which does not comply with Ohm's law (Fig. 8.3). The nonlinearity of this characteristic is displayed sharper with increasing voltage V_d. When V_d is further increased the voltage across the most narrow portion (near the drain) of the channel reaches the value of V_0 and the channel between the *p-n* junctions becomes almost pinched off (Fig. 8.2*d*). The transistor thus reaches a state at which an increase in the drain voltage leads to a sharp rise in the differential resistance impeding further increase of the current and consequently the drain current becomes practically invariable. For this reason the current-voltage characteristic has a portion of saturation of the drain current (Fig. 8.3). With a further increase of the voltage between the source and drain the field strength in the space charge region can rise so intensely that it may cause breakdown in the drain-gate circuit.

Thus, unlike the regular transistors controlled by the input current (emitter current), the field-effect transistors are governed by the input voltage (gate voltage).

A distinguishing feature of the field-effect transistor is the absence of injection of minority carriers across the *p-n* junction. The transport of current in this type of transistors is effected by the majority carriers of a given semiconductor. This is why the field-effect transistors are also called unipolar. It should be remembered that in common transistors, termed bipolar, the carriers of current are both majority carriers in the emitter and collector regions and minority carriers in the base region.

Apart from the cutoff voltage V_0 mentioned above, the field-effect transistors are characterized by the following main parameters: maximum drain current $I_{d\ max}$; input resistance R_{in}; interelectrode capacitances (gate-source C_{gs}, gate-drain C_{gd}, drain-source C_{ds}); maximum working frequency f_{max}; and transadmittance S which characterizes amplification properties of a transistor and represents a change in the drain current with an increment of the gate voltage by one volt at a constant drain voltage.

Field-effect *p-n* junction transistors are made mainly by fusion and diffusion.

Germanium *p-n* junction field-effect transistors, prepared by the alloying method have the following values of typical parameters: $V_0 = 10$ to 30 V; $I_{d\ max} = 5$ to 30 mA; $S_{max} = 2$ to 5 mA/V; $C_{gs} \approx C_{gd} = 30$ to 40 pF; $f_{max} = 200$ to 500 kHz.

Better values are impossible to obtain as the alloying technique does not allow the control of the depth of penetration and consequently the geometry of the channel to an accuracy better than several microns, and, besides, this technique is impracticable in preparing small-area gates.

The sequential diffusion method permits of adjusting the thickness of diffused layers with a precision of tenths fractions of a micron. What is more, the planar technique and the use of special masks help attain more advantageous relationships between the dimensions of active regions in field-effect devices.

The parameters of field-effect transistors with control *p-n* junctions made by planar methods are considerably better than those made by alloying: $V_0 = 0.8$ to 7 V; $I_{d\ max} = 1$ to 15 mA; $S_{max} = 1$ to 3 mA/V; $C_{gs} \approx C_{gd} = 2$ to 10 pF; $f_{max} = 10$ to 70 MHz. This accounts for their large-scale production and wider use.

The planar technique combined with the methods of sequential epitaxial growing can obviously improve still further the parameters of the field-effect transistors.

8.2. INSULATED-GATE FIELD-EFFECT TRANSISTOR

In conventional *p-n* junction field-effect transistors the current is controlled by voltage variations across the gate that constitutes a region whose conductivity type is opposite to that of the channel region. Another method of control of the channel resistivity is by changing the potential across an electrode which is also called gate, but insulated from the semiconductor bulk.

The transistors operating on this principle are referred to as insulated-gate field-effect transistors or MIS transistors, i.e., metal-insulator-semiconductor transistors. Since in most cases oxide, such as SiO_2, is used as an insulator, these devices are also termed MOS transistors (metal-oxide-semiconductor).

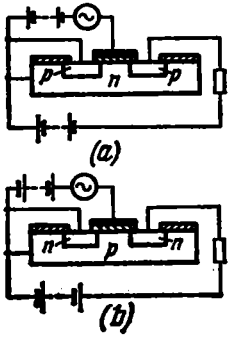


Fig. 8.4. MIS transistors with channels of different types
(*a*) *p*-channel transistor; (*b*) *n*-channel transistor

The potential difference developed between the semiconductor and the insulated electrode (gate) results in the creation near the semiconductor surface of a layer of carriers whose density differs from the density of carriers in the semiconductor bulk. For example, a positive potential impressed on the control electrode adds to the density of carriers of the same sign (enrichment region) in the *n*-type semiconductor. By varying the voltage across the gate so as to increase the carrier density in the channel, it is possible to build up the current through the channel. At a sufficiently large voltage of opposite sign (barrier voltage) the current through the channel can also rise due to the appearance of an inversion layer at the surface.

Thus, in the *n*-type semiconductor a positive voltage across the gate causes accumulation and a negative voltage, depletion of carriers at the surface, while a considerably large negative voltage leads to surface inversion. In the *p*-type semiconductor a similar process of the development of an *n*-type inversion layer is observed on applying a positive voltage between the gate and the semiconductor bulk, called substrate.

The inversion layer thickness (Debye length) and the carrier density in the layer are entirely dependent on the nature of a semiconductor and the magnitude of the voltage impressed across the gate.

The *n*-channel insulated-gate field-effect transistor represents a structure made up of two highly doped *n*-regions (source and drain) arranged in a *p*-semiconductor and spaced at a small distance from each other, and also of a metal electrode (gate) insulated from the semiconductor (Fig. 8.4*a*). The positive voltage applied to the gate tends to induce a channel between the *n*-regions, i.e., an inversion *n*-layer. By controlling the conduction of this layer it is possible to vary the current through it and thus amplify the power of the signal applied to the gate. There is practically no drain current in this transistor when the bias is not applied on the gate. Such a transistor is called induced channel-type transistor.

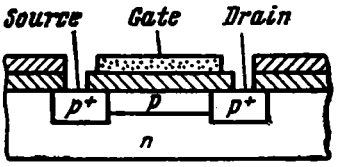


Fig. 8.5. MIS transistor with a built-in channel of *p*-type

The *p*-channel metal-insulator-semiconductor transistor structure is shown in Fig. 8.4*b*. The induced current-conducting channel in the MIS transistor is formed only under certain polarity and at a voltage across the gate which exceeds a certain threshold voltage V_{th} sufficient to bring about the inversion of a near-surface layer (see Ch. One). For this reason they operate only in the enrichment mode when increasing voltage across the gate causes the conduction of the channel to grow, raising thereby the drain current.

A variety of the MIS structure is a device in which the channel between the source and the drain is made preliminarily by certain methods (Fig. 8.5). Such a transistor has a channel between the source and the drain when there is no voltage across the gate, and therefore this transistor is called the build-in-channel transistor as distinct from the induced channel-type transistor. The carrier density in the channel can drop off or rise depending on the polarity of the voltage applied. This transistor operates both in the enhancement and the depletion mode.

Now we consider the operation of the induced-channel MIS transistor in more detail. Let us denote the channel length by l, channel width by d and insulator thickness by d_0 (Fig. 8.6).

The channel conductance at some arbitrarily chosen point x may be considered as a function of the voltage drop, $V(x)$, at this point and of the gate voltage V_g. The potential of the source is taken to equal zero (the source is grounded). Assuming that most of the external voltage applied to the gate drops across the insulator layer, the

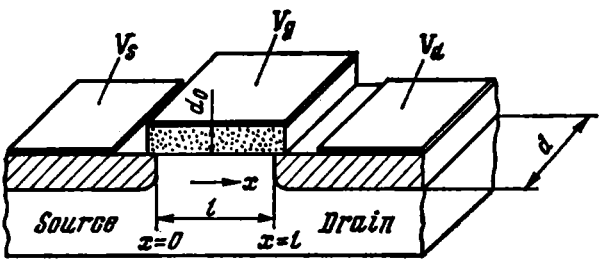


Fig. 8.6. MIS transistor structure

transverse field strength in the insulator layer is

$$E_y(x) = \frac{V_g - V(x)}{d_0} \tag{8.2}$$

The density $\rho(x)$ of the space charge induced in the semiconductor per unit area of the gate is defined as

$$\rho(x) = \varepsilon_{in}\varepsilon_0 E_y(x) = \frac{\varepsilon_0\varepsilon_{in}}{d_0}[V_g - V(x)] \tag{8.3}$$

where ε_{in} is the relative permittivity of the gate insulator.

As the voltage difference between the gate and the channel does not reach V_{th} the channel is not yet induced and the density of the mobile carrier charge in the channel is

$$\rho(x) = 0 \text{ when } V_g - V(x) \leqslant V_{th} \tag{8.4}$$

At higher voltages

$$\rho(x) = \frac{\varepsilon_{in}\varepsilon_0}{d_0}[V_g - V(x) - V_{th}] \quad \text{when}$$
$$V_g - V(x) > V_{th} \tag{8.4a}$$

The conductance of the channel portion of length dx is

$$G(x) = \frac{\rho(x)\mu d}{dx} \tag{8.5}$$

The density of the drain current I_d can therefore be determined from the expression

$$I_d = G(x)\,dV = \frac{\varepsilon_{in}\varepsilon_0\mu d}{d_0}[V_g - V_{th} - V(x)]\frac{dV}{dx} \tag{8.6}$$

where dV is the voltage drop across the portion dx.

Since the same drain current I_d passes through the channel section of any size, in other words, as I_d does not depend on the coordinate x, then integrating (8.6), we can easily obtain

$$I_d = \frac{\varepsilon_{in}\varepsilon_0\mu d}{d_0 l}\left[(V_g - V_{th})V_d - \frac{V_d^2}{2}\right] \tag{8.7}$$

The expression thus derived holds on condition that $V_g - V_d > V_0$.

When $V_d \geqslant V_g - V_0$ the drain current is brought to saturation and remains almost constant with any further increase of V_d. The saturation current is determined by substituting $V_d = V_g - V_0$ into expression (8.7)

$$I_{ds} = \frac{\varepsilon_{in}\varepsilon_0\mu d}{2d_0 l}(V_g - V_0)^2 \tag{8.8}$$

Transadmittance $I_d(V_g)$ on the I/V curve in the saturation region is determined from expression (8.8)

$$S_s = \frac{dI_d}{dV_g}\bigg|_{V_d = V_g - V_0} = \frac{\varepsilon_{in}\varepsilon_0\mu d}{d_0 l}(V_g - V_0) \tag{8.9}$$

The frequency properties of MIS transistors are characterized mainly by the charge time constant of the capacitance formed between the gate and the semiconductor bulk across the channel resistance.

The minimum resistance of the completely opened channel is nothing but $1/S_s$. Hence, the cutoff frequency f_{max} may be expressed as

$$f_{max} = \frac{1}{2\pi\tau} = \frac{S_s}{2\pi C_g} \tag{8.10}$$

where τ is the charge time constant of the gate capacitance and C_g is the capacitance of the gate relative to the source and the drain.

The gate may be roughly regarded as a parallel-plate capacitor with plates of length l and width d spaced apart at distance d_0, the space being filled with the dielectric of permittivity $\varepsilon_{in}\varepsilon_0$.

Hence,

$$C_g = \frac{\varepsilon_{in}\varepsilon_0 ld}{d_0}$$

Substituting the value of S_s from (8.9) into (8.10) yields

$$f_{max} = \frac{\mu_n}{2\pi l^2}(V_g - V_0) \tag{8.11}$$

The main material used for the fabrication of insulated-gate field-effect transistors is silicon. Silicon field-effect transistors with silicon dioxide insulators are produced by conventional planar technique. The production process involves careful control over the geometrical dimensions of the structure and the thickness of the layers, which is prerequisite for obtaining required parameters with a small scatter in their values.

A decrease in the thickness d_0 of the insulator layer causes an increase in the transadmittance and a drop in the threshold voltage though leads at the same time to an undesirable rise of the gate capacitance. An increase in the channel width results in an increase of not only $I_{d\ max}$ and S but also C_g.

MIS transistors' parameters attained nowadays are sufficiently good: f_{max} = 300 to 400 MHz; V_{th} = 2 to 10 V; C_{in} = 0.5 to 3pF; S_{max} = 0.5 to 4 mA/V; $R_{in} = 10^{10}$ to 10^{15} ohms.

The speed of response of MIS transistors can be enhanced still further by diminishing their channel length.

The channel length can be brought down to 5-6 μm by the regular diffusion methods, though in this case the triode exhibits a considerably high gate-drain parasitic capacitance limiting the cutoff frequency. This capacitance is determined by the extent of pinch-off of the gate and the drain (and also of the source) and reaches 2 to 3 pF.

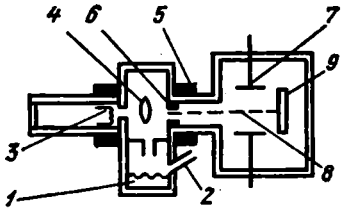


Fig. 8.7. Arrangement for ion implantation of impurities into a semiconductor

1—ion source; *2*—vent for loading impurities to be evaporated; *3*—ionizing electron gun; *4*—ion plasma; *5*—coils inducing a magnetic field; *6*—accelerating electrodes; *7*—deflecting electrodes; *8*—ion beam; *9*—substrate

To obviate the pinch-off of the gate and drain regions use is made of the ion implantation method by which the impurities are introduced in a semiconductor by an ion beam and not by diffusion from the gaseous phase. This method is based on irradiation of a single crystal of silicon by high-energy ions. By bombarding the *n*-type silicon with aluminium or boron ions it is possible to obtain the regions of *p*-type conductivity, while bombardment of the *p*-type silicon with phosphorus ions results in the formation of *n*-type regions.

The silicon dioxide in this case ceases to be a masking agent which permits carrying out selective diffusion from the gaseous phase, since the SiO_2 proves practically transparent to high-energy ions.

In the ion implantation technique, silicon is coated with an oxide layer, then gate electrodes are deposited and the wafer is irradiated by high-energy ions.

The metal electrode completely absorbs the ions, while the unprotected portions allow the impurity ions to penetrate into the bulk of the crystal and form a required conductivity type. The depth of penetration of the ions into silicon is very small. Phosphorus and boron ions with an energy of 100 keV penetrate to a depth of 0.4 and 0.6 μm, respectively, while with an increase of the ion energy to 1 MeV the depth of penetration of ions increases merely to 1.5-2 μm.

Owing to this method the pinch-off of the gate regions with respect to the source and the drain does not occur in principle, and therefore the speed of response of MIS transistors sharply increases. The MIS transistors made by the ion implantation method operate at a frequency in excess of 1GHz.

The main unit in the ion implantation setup is an ion gun. The general arrangement of the setup is shown in Fig. 8.7. Phosphorus, boron or aluminium are loaded into evaporator *1*. Vapours are ionized in ionization chamber *4* under the action of a high electric or magnetic field or an electron beam emitted by electron gun *3*. The ion beam is focused onto the crystal surface by the magnetic *5* and electric *7* control system.

Another way of lowering the parasitic capacitance is the fabrication of MIS transistors with silicon gates.

For this purpose the silicon surface is oxidized and coated with a polycrystalline silicon layer (Fig. 8.8*a*) which is then subjected to photoengraving (Fig. 8.8*b*) to shape it to the configuration suitable for the deposition of gate electrodes.

Afterwards the oxide covering the regions to be used for the source and the drain is removed by the etch (hydrofluoric acid HF) which does not affect silicon. The silicon film serves as a mask for protecting the channel region, and the gate is brought here into coincidence with the channel automatically (Fig. 8.8*c*).

The next step includes simultaneous alloying of the source, drain and gate regions. The gate so obtained has the same type of conductivity and practically the same carrier density as the source and the drain regions (Fig. 8.8*d*). This production process helps reduce the threshold voltage of MIS transistors to 1.8 V, widening thereby the field of application of these transistors.

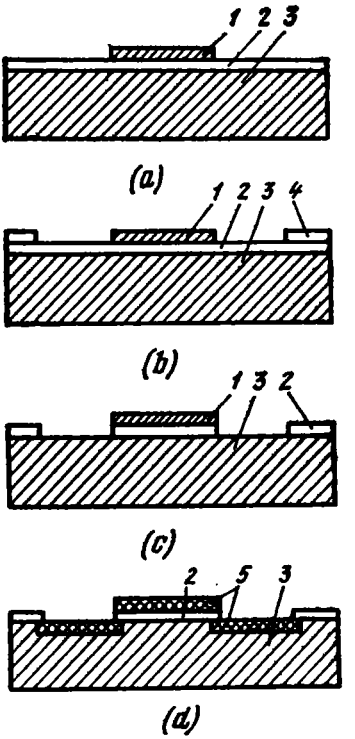


Fig. 8.8. Steps of preparation of the MIS transistor with a silicon gate

1—silicon film; *2*—oxide layer; *3*—*n*-type substrate; *4*—photoresist; *5*—*p*-silicon deposited from the gaseous phase

The silicon gate makes it possible to lower the silicon capacitance to 0.1-0.05 pF and sharply increase the speed of response and, above all, raise the density of packing of transistors on the crystal surface since the coincidence of the gates with the channel is achieved here automatically (their noncoincidence is one of the main causes of rejects of MIS transistors).

In production of MIS transistors, such materials as CdS, ZnS, GaAs and Ge are also employed.

Use is generally made in practice of field-effect transistors in a common-source, comon-drain and common-gate connection.

The common-source connection allows for voltage amplification and has a high input and output impedance. This connection diagram is comparable to that of an electron tube in a common cathode connection, the transistor gate being analogous to the control grid (Fig. 8.9*a*).

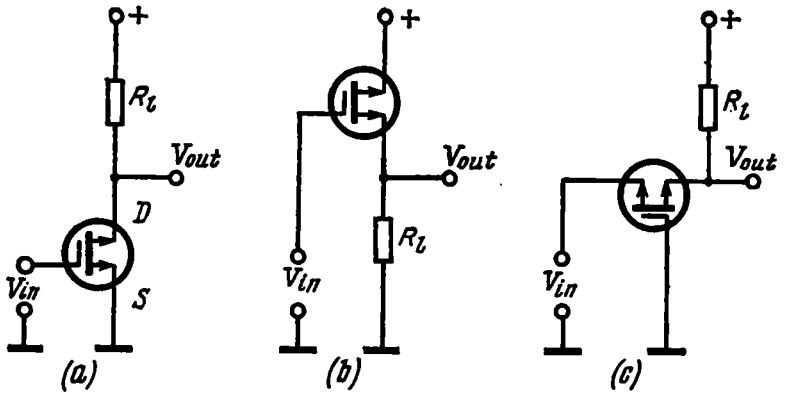


Fig. 8.9. MIS transistor configurations
(a) common source; (b) common drain; (c) common gate

The output signal in this circuit is opposite in phase to the input signal. Since by formula (8.9) the transadmittance is a function of the gate voltage, the amplification factor is heavily dependent on the mode of operation.

The common-drain circuit configuration (Fig. 8.9*b*) resembles that of a vacuum-tube cathode follower. The voltage gain factor of such a circuit is little dependent on the magnitude of the signal. The input signal is in phase with the output signal and the input impedance is higher and the output impedance is lower than those in the common-source circuit. For this reason this circuit is used for matching pulse sources with high output impedance and stages with low input impedance.

The common-gate connection (Fig. 8.9*c*) is similar to that of an electron tube in a common-grid connection.

The advantages of MIS transistors over bibolar transistors are:

(a) high input impedance which depends only on the leakage resistance of the insulator. It runs as high as 10^{12} to 10^{15} ohms as measured by the direct current;

(b) low noise level, of the order of 0.5 db, which is attributed to a small effect of recombination processes since in the field-effect transistors, as opposed to the bipolar transistors, only majority carriers are allowed to conduct current;

(c) relative ease of fabrication.

Field-effect transistors have found extensive use in various amplifying and switching circuits. They are often employed in combination with bipolar transistors. Owing to their high impedance, the field-effect transistors can be used for matching circuits composed of bipolar transistors with high-resistivity transducers. A low noise level

of the field-effect transistors, reaching only 0.5 to 1 db, allows their advantageous use in input stages of low-noise amplifiers.

Switching circuits made up of unipolar transistors offer certain advantages over bipolar transistor circuits. First, the switching circuits based on unipolar transistors are slightly simpler; second, unipolar transistors can be used to construct devices with low power consumption and very small working currents; and third, they can be used in direct-coupling circuits of large load capacity.

The limitation of field-effect transistors is their relatively high resistance in the operational condition, which is by an order of magnitude higher than the resistance of bipolar transistors.

The use of MIS transistors in switching circuits is limited due to a noticeable instability of their characteristics, the cause of which is the movement of charges formed by vacancies and ions of impurities (specifically alkali metals) in the oxide under the action of an electric field. Something of this kind was observed in planar bipolar transistors at early stages of their development, but in MIS transistors this effect is more pronounced.

8.3. UNIJUNCTION TRANSISTOR

The unijunction transistor, more often called the double-base junction diode, is a semiconductor structure with one *p-n* junction and three terminals. Its current-voltage characteristic has a negative resistance region and its principle of operation is based on variations in the bulk resistance of the semiconductor in response to the injection of carriers. The schematic arrangement of a double-base diode and its connection diagram are shown in Fig. 8.10*a*.

The double-base diode is a single-crystal semiconductor with ohmic contacts at each end and a *p-n* junction (injector) between them. The voltage is so applied to the ohmic contacts (bases) that the current flowing between them causes a voltage drop in the semiconductor bulk under the *p-n* junction and thus cuts off the junction.

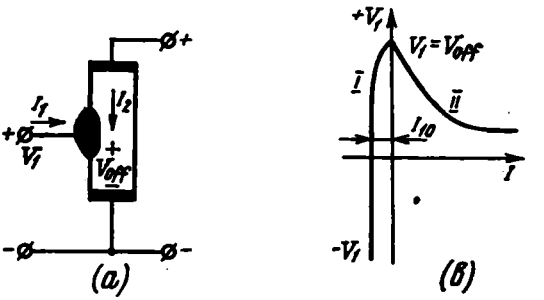


Fig. 8.10. Unijunction transistor (*a*) and its characteristic (*b*)

A very small and practically constant reverse current I_{10} flows through the blocked *p-n* junction. This state of the device is represented by portion *I* on the current-voltage characteristic (Fig. 8.10*b*). For the *p-n* junction to be made conductive, it is necessary that the voltage across the injector should exceed the blocking voltage drop at some point in the base under the injector. At a sufficiently increased injector voltage the *p-n* junction becomes conducting at this point.

The injected carriers reduce the intrinsic resistance of the base and the voltage drop under the junction decreases, leading thus to still larger triggering of the *p-n* junction. This gradually accelerating process is reflected on the current-voltage characteristic by portion *II* with a region of negative differential resistance.

Two-base diodes are mainly used in relaxation circuits (multivibrators, pulse counters). In order to devise such a circuit it is often sufficient to employ only one diode. The shortcomings of the two-base diodes are a slow switching speed and large consumption of energy. These disadvantages are responsible for a rather limited production and application of the two-base diodes.

8.4. *p-n-p-n* SWITCH

The *p-n-p-n* switch is a semiconductor device with a four layer structure and three *p-n* junctions. Depending on the number of leads brought out from the switch one can differentiate a two-terminal device (dinistor), a three-terminal device (trinistor or thyristor), and a four-terminal device (binistor). All these devices have a negative resistance region on the current-voltage characteristic.

Consider now the principle of operation and the parameters of a *dinistor* (Fig. 8.11*a*). Let us apply the voltage to the device so that the outer *p-n* junctions can be cut off. A very small current will flow through the device, plotted as a reverse branch on the current-voltage curve (Fig. 8.12).

Now we reverse the polarity of the voltage applied. Junctions *1* and *3* will become biased in the forward direction and act as emitters in respect to the inner *n*- and *p*-regions of the device. Junction *2* will be rendered nonconductive and therefore, disregarding a very small voltage drop across the bulk of the semiconductor and that across the forward-biased *p-n* junctions, we may assume that all the voltage is impressed on junction *2*. A small current will still continue to flow through the device (region *I* on the current-voltage curve of Fig. 8.12).

The dinistor may thus be regarded as a system of superposed *n-p-n* and *p-n-p* transistors (Fig. 8.11*b*) having common regions. Designating the current gain factors of these transistors as α_1 and α_2, we find the current flowing through the reverse-biased junction *2*:

$$I_2 = \alpha_1 I_1 + \alpha_2 I_3 + I_{rev} \tag{8.12}$$

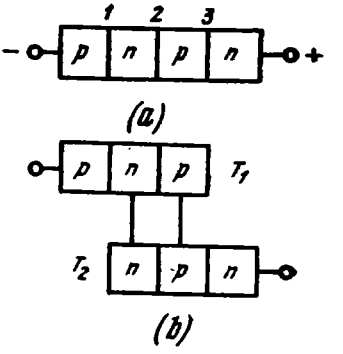


Fig. 8.11. Dinistor structure (*a*) and its two-transistor representation (*b*)

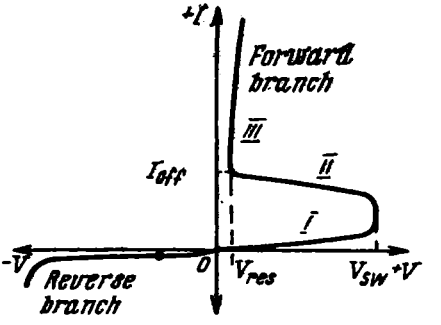


Fig. 8.12. Current-voltage characteristic of a dinistor

where I_1, I_2 and I_3 are currents through respective *p-n* junctions, and I_{rev} is the reverse current of the reverse-biased device.

Since all the junctions are connected in series,

$$I_1 = I_2 = I_3 = I \tag{8.13}$$

It thus follows from (8.12) that

$$I = \frac{I_{rev}}{1-(\alpha_1+\alpha_2)} \tag{8.14}$$

Further rise in the applied voltage causes a decrease in the potential barrier of the reverse-biased *p-n* junction followed by a redistribution of the voltage drop throughout the entire structure, reduction of the potential barriers of the *p-n* junctions and an avalanche-like increase in the number of injected carriers. The carrier multiplication process starts playing a prominent role. This process is represented by a negative resistance region *II* on the current-voltage curve. As soon as the dinistor is brought to the forward-bias state, its intrinsic resistance becomes small and further increase of the voltage results in a sharp rise of the current (region *III* on the current-voltage curve). In this region the voltage, called the residual

voltage of the dinistor in the saturated mode of operation, is little dependent on the current. The condition of transition of the dinistor from the turn-off to the turn-on state may be found from the analysis of expression (8.14).

Indeed, when $\alpha_1 + \alpha_2 = 1$ the denominator of (8.14) goes to zero, which means that at a certain voltage the current will rise sharply and the dinistor will leave the region of positive resistance $R = \frac{dV}{dI} > 0$ and enter the region of negative resistance $R = \frac{dV}{dI} < 0$, as actually observed on portion *II* of the current-voltage characteristic of the dinistor.

For the dinistor to remain in the open state, the value of the current through the dinistor must be higher than the value of the turn-off current I_{off}.

The dinistor operates as a switch in a driven mode, staying in the off condition as long as the applied voltage remains lower than the switching voltage V_{sw}. Sometimes bias E_b is applied on the dinistor so as to render it conductive by a lower voltage as compared to V_{sw}. The dinistor becomes open when condition $E_b + V > V_{sw}$ is fulfilled.

The time characteristic describing the switching properties of the dinistor is the value of the turn-on time t_{on} and turn-off time t_{off}. The turn-on time is the time interval from the moment the trigger pulse has been applied to the dinistor to the moment when the value of the current through the dinistor reaches 0.9 of the maximum required value. The turn-off time is the time interval that the dinistor takes to pass into the off condition from the moment the blocking pulse has been applied to it.

The t_{off} is always larger than the t_{on}, which is attributed to the fact that after cutting off the supply of the turn-on voltage a fall in the current is delayed somewhat due to the storage of excess carriers. This process is suggestive of the process of recovery of the diode reverse resistance and proceeds faster, when the lifetimes of the minority carriers within the bulk of the dinistor is lower.

Dinistors are generally used for two-position switching under the action of control pulses.

Owing to the negative resistance region displayed by its current-voltage characteristic, the dinistor can serve as a relaxation oscillator.

The thyristor (or *trinistor*) is a semiconductor device consisting of a four-layer *p-n-p-n* structure having terminals brought out from two extreme regions and from one of the inner regions (Fig. 8.13). The thyristor differs from the dinistor by the number of terminals (three as against two in the latter). A significant advantage of the thyristor over the uncontrolled diode, i.e. dinistor, is the possibility of controlling the thyristor switching voltage by applying a voltage to

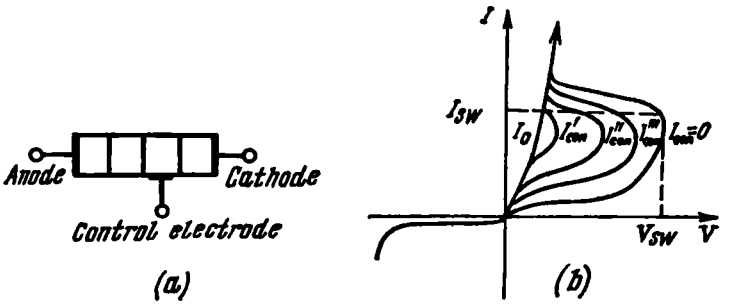


Fig. 8.13. Thyristor (*a*) and its current-voltage characteristics (*b*)

the third (control) electrode. This is why the thyristors, in distinction to the dinistors, are sometimes called controlled switches.

The control electrode may be connected to any of the two inner regions (bases) of the thyristor. For this purpose, the polarity of the control voltage must be changed in conformity with the polarity of connection of the emitter junction adjacent to the base.

As in the case of the dinistor, the thyristor structure may be presented in the form of two transistors, with the switching taking place when $\alpha_1 + \alpha_2 = 1$.

The current-voltage characteristic has the same specific regions, though here it is manifested as a family of curves whose parameter is the current passing through the control electrode (Fig. 8.11*b*). The outer curve reflects the condition when $I_{con} = 0$, i.e., it is entirely identical with the dinistor characteristic.

The higher the control electrode current, the smaller the voltage at which the thyristor becomes biased in the on condition. Starting with a certain value of I_{con}, called the current of curve rectification, the current-voltage characteristic takes the form which is similar to the forward current-voltage curve of the common diode.

Thyristors are used in switching large currents from hundreds of milliamperes to hundreds of amperes. While passing through the thyristor, these currents cause intensive power dissipation attended with noticeable heating of the device despite the fact that special care is taken to improve the heat withdrawal. The effect of temperature on thyristor parameters must therefore be given particular consideration.

Rising temperature is responsible for a sharp rise of the current in the off condition, increase of the values of α_1 and α_2 and increase of the turn-off time since the lifetimes of minority carriers become longer, and also a decrease in the turn-on current of the control electrode and the turn-off current.

Table 8.1

Electric Parameters of Dinistors and Thyristors

Parameter	Symbol	Unit of measurement	Dinistor type								Thyristor type			
			228А	228Б	228В	228Г	228Д	228Е	228Ж	228И	235А-Г	235А-Е	УПВК	ВКДУ
Switching voltage	V_{sw}	V	10-20	14-28	20-40	28-56	40-80	56-112	80-160	100-200	40-100	50-150	150-1500	50-700
Residual voltage, not more than	V_{res}	V	1.5	1.5	1.5	1.5	1.5	1.5	1.5	1.5	2	2	0.5-0.8	0.9-1.25
Reverse voltage, not less than	V_{rev}	V	—	—	—	—	—	—	—	—	50	50-150	100-1200	50-700
Control current, not more than	I_{con}	mA	—	—	—	—	—	—	—	—	20-50	150	150	1000
Maximum current	I_{max}	A	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	2	10	50-100	50-200
On time, not more than	t_{on}	μs	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	5	10	2-5	10
Off time, not more than	t_{off}	μs	5	5	5	5	5	5	5	5	35	35	15-20	25
Turn-off current, not more than	I_{off}	mA	15	15	15	15	15	15	15	15	—	—	—	—
Reverse current, not more than	I_{rev}	mA	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	—	—	—	—
Switching current, not more than	I_{on}	mA	1	1	1	1	1	1	1	1	—	—	—	—

With the rise in α_1 and α_2, at elevated temperatures the condition $\alpha_1 + \alpha_2 = 1$ begins to hold at smaller voltages, which leads not only to the narrowing of the working temperature range of the thyristor but also to occurrences of arbitrary switching.

Various methods are used to prevent a spontaneous action of the device. One of the widespread circuit modification solutions consists in applying a certain back bias on the control electrode. A disadvantage of this method is a decrease in the responsiveness of the thyristor and the need for a corresponding increase of the control signal. A design solution of the problem involves the connection of one of the emitter junctions with the adjacent base, deliberately limiting therewith the value of α and thus affording switching at higher voltages.

The electric parameters of some dinistors and thyristors made in the USSR are given in Table 8.1.

The switching device with a *p-n-p-n* structure and four terminals brought out from all the four layers is called *binistor*. Its current-voltage characteristic resembles in general those of the thyristor and the dinistor as they are in essence special cases of the binistor, differing from the binistor only in that the thyristor lacks one control electrode and the dinistor, two electrodes.

The capabilities of the binistor are more extensive as the additional control electrode allows one not only to shift the current-voltage characteristic but also to change its shape to certain limits.

Among the multilayer-structure devices, it is the thyristor that has found most expanding applications. Thyristors are mainly used in switching and stepless voltage-regulation circuits of medium- and high-power radio and electrical devices.

Review Questions

1. Explain the principal difference between the bipolar and unipolar transistors.
2. Name the varieties of field-effect transistors.
3. Draw three connection diagrams of MIS transistors.
4. Plot and explain the current-voltage characteristic of a two-base diode.
5. Name the versions of *p-n-p-n* switches and explain the mechanism of switching.
6. What is the effect of temperature on the main parameters of the thyristor?

Chapter Nine

THERMOELECTRIC AND GALVANOMAGNETIC DEVICES

9.1. THERMOELECTRIC PHENOMENA IN SEMICONDUCTORS

As far back as 1821 Seebeck discovered that at the ends of the circuit composed of dissimilar conductors there appears an electromotive force if the junctions of the conductors are held at different temperatures. The value of this thermal emf depends on the temperature difference between the cold and the hot junction and may be expressed as

$$E = \alpha_T \Delta T$$

where α_T is the thermoelectric coefficient expressed as the potential difference per degree C change in temperature between the cold and the hot junction.

Metals, experiments on which helped discover thermoelectricity, have a very small value of α_T that does not exceed tens of microvolts per degree C, and, as a result, the Seebeck effect had long been utilized only in thermocouples. Thermal emfs of different metals are taken with reference to lead. The values of emfs for some metals are given in Appendix 4.

Advances in semiconductor technology made it possible to use the Seebeck effect for the generation of energy, i.e., for direct conversion of heat into electricity, owing to the fact that the thermoelectric coefficients of semiconductors are considerably higher than those of metals. For example, the thermoelectric coefficient of cuprous oxide reaches 1200 μV/°C.

Let us consider the mechanism of generation of the thermal emf in an electron semiconductor. As soon as the temperature gradient is developed in the semiconductor, the density of electrons at the hot end rises and so does their kinetic energy, and the electrons begin to diffuse from the hot region into the cold region of the semiconductor. The flow of electrons from the hot region results in the formation of an excess positive charge at the hot end and of a negative charge at the cold end. Thus, opposite charges appear at the ends of the semiconductor. The counterflow of electrons at the initial moment is by far smaller than the flow of electrons from the hot end.

The space charges of opposite signs that have appeared at the ends of the semiconductor bring about the development of an internal field in the semiconductor, which impedes further flow of electrons from the hot to the cold region. The field finally equalizes the flows of electrons. The electromotive force established under this condition will be referred to as the semiconductor thermal emf at a given temperature.

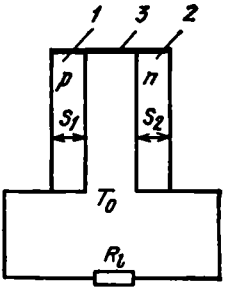


Fig. 9.1. Diagram of a thermoelectric circuit made up of a positive (*1*) and a negative (*2*) arm interconnected by electrode (*3*)

The sign of the coefficient α_T depends on the type of semiconductor. In the n-type semiconductor the hot end has a positive and the cold end, a negative charge. In the p-type semiconductor the hot end corresponds to a negative pole. The thermal emf in the electron semiconductor has thus a minus sign and that in the hole semiconductor, a plus sign.

The value of α_T in semiconductors having both donor and acceptor impurities varies with the density and mobility of electrons and holes and as a rule is significantly smaller than that in purely electron or hole semiconductors. If the density and mobility of electrons are respectively equal to the density and mobility of holes, the thermal emf in such a semiconductor is zero.

The quantity α_T is related to the density and effective mass m^* of charge carriers of the same type by the Pisarenko formula

$$\alpha_T = \pm \frac{k}{q} \left[B + \ln \frac{2\,(2\pi m^* kT)^{3/2}}{nh^3} \right] \tag{9.1}$$

where B is the constant which depends on the type of crystal lattice: $B = 2$ for an atomic lattice and $B = 3$ for an ionic lattice.

From the analysis of formula (9.1), it becomes clear why the value of α_T for metals is so small. As the density and energy of electrons in metals depend rather little on temperature, the flow of electrons from the hot to the cold end is insignificant.

Let us now form a series circuit from an n-type and a p-type semiconductor (Fig. 9.1) and start heating the contact (junction) between them. The cold end of the hole semiconductor will charge positively and that of the electron semiconductor, negatively with respect to the hot junction, i.e., the potential difference at the cold ends will be the sum of the thermal electromotive forces in the electron and the hole semiconductor.

Another thermoelectric phenomenon that has recently found application over a hundred years since it was first discovered by Peltier in 1834 is the Peltier effect which shows up as the production or absorption (depending on the direction of current) of a certain amo-

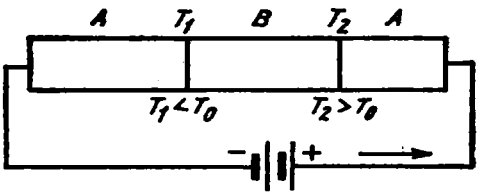


Fig. 9.2. Temperature gradient developed at the contact between dissimilar elements

unt of heat Q at the junction between two conductors when a current is passed through it, the heat being proportional to the current through the junction:

$$Q_P = PI \tag{9.2}$$

where P is the Peltier coefficient.

The Peltier effect, weakly pronounced in metals, was considered for a long time as a curious phenomenon having no technical importance (let us recall that the Seebeck effect was utilized at that time in thermocouples).

In the course of time, with the development of semiconductor technology the Peltier effect has found use in thermoelectric coolers.

The Peltier effect results from the difference in the kinetic energy of electrons between dissimilar conductors. If the direction of current is such that higher-energy electrons move to the conductor with lower-energy electrons, heat is produced at the junction because the excess energy of electrons is transferred to the crystalline lattice.

As soon as the direction of current is reversed, the lower-energy electrons will pass to the conductor with the higher-energy electrons and the junction will begin to cool down since the incoming electrons must replenish their energy at the expense of the energy of the conductor lattice. If we compose a circuit from different elements in such a manner (Fig. 9.2) that the kinetic energy of carriers in branch A can be higher than that in branch B, and also pass the current in the direction shown in the Figure, a temperature difference will appear between the junctions AB and BA:

$$\Delta T = T_2 - T_1$$

Reversing the current direction results in the change of the sign of the temperature difference.

The relation between the Seebeck and the Peltier coefficient is given by

$$P = \alpha_T T$$

The Peltier effect shows up most vividly at the boundary between a p-type and an n-type semiconductor (Fig. 9.3). In this case the difference in the kinetic energy of carriers on both sides of the junction is particularly great. When electrons move from the n-type

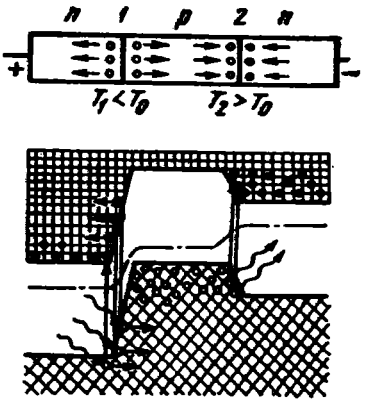


Fig. 9.3. Energy-band diagram of a semiconductor-metal contact, illustrating the development of the Peltier effect

to the *p*-type semiconductor, they recombine with holes at junction *2* (Fig. 9.3). During this process the electrons are transferred from the conduction band to the valence band and the energy of interaction between the electrons and holes is transformed into heat which is liberated at the junction, and the temperature of the junction thus rises over the ambient temperature T_0, in other words, T_2 exceeds T_0.

The electrons in the *n*-region and holes in the *p*-region are drawn away from junction *1*. As this takes place, the electrons from the valence band of the *p*-semiconductor are pulled over by the *p-n* junction field to the conduction band of the *n*-semiconductor. The electrons and holes being produced in the process of thermal generation move away from the junction, and the energy ΔE expended on the generation of electron-hole pairs is compensated for by the lattice energy, with the result that the junction cools down.

Apart from the Seebeck and Peltier effects, the Thomson effect also belongs to the class of thermoelectric phenomena. The Thomson effect is manifested as the generation or absorption of heat when a current is passed through a conductor with a temperature gradient, the amount of heat being in proportion to the current and the temperature gradient:

$$\Delta Q_\tau = \tau I \frac{dT}{dx} \tag{9.3}$$

The physical essence of the Thomson phenomenon is the heating of a semiconductor when electrons move from the hot end to the cold one and its cooling when electrons move from the cold to the hot end. The Thomson coefficient is related to the thermoelectric coefficient by

$$\tau = T \frac{d\alpha_T}{dT} \tag{9.4}$$

The amount of heat produced as a result of the Peltier and the Thomson effects varies directly with the current through the junction, as is obvious from formulas (9.2) and (9.3). In addition, when the current is passed through a semiconductor, the latter releases the Joule heat which increases as the square of current I^2. This is why thermoelectric phenomena are observed only at small currents since with higher currents the Joule heat begins to play a dominant role independent of the direction of current and against the background of this process the Peltier and Thomson effects may be unnoticeable at all.

9.2. THERMOELECTRIC GENERATORS

We consider now a thermoelectric circuit made up of two bars of n-type and p-type semiconductors (Fig. 9.1). The emf of this circuit is the sum of the thermal emfs in the hole and the electron portion of the circuit:

$$\alpha_T = \alpha_{Tn} + \alpha_{Tp}$$

According to the energy conservation law, the power lost in the resistive load R_l must be balanced out by the difference between thermal flows, one being delivered to the hot end and the other dissipated at the cold end. Hence,

$$P_l = I^2 R_l = Q_1 - Q_2 \tag{9.5}$$

where Q_1 is the heat delivered to the hot junction, and Q_2 is the heat given up at the cold junction.

The thermal flow generated in a thermoelement when current is passed through it is made up of the heat produced by the Joule effect (Joule heat Q_J) and Peltier effect (Peltier heat Q_P) plus the heat flow transferred from the hot to the cold junction due to the thermal conduction of the thermoelement. Let us now examine each of these thermal flows.

The heat liberated at the hot junction as a result of the Peltier effect may be written in the form

$$Q_{P1} = (\alpha_{Tn} + \alpha_{Tp}) I T_1 \tag{9.6}$$

and at the cold junction, as

$$Q_{P2} = (\alpha_{Tn} + \alpha_{Tp}) I T_0 \tag{9.7}$$

The thermal flow transferred from the hot junction to the cold one due to heat conduction is

$$Q_\varkappa = (\varkappa_n S_n + \varkappa_p S_p)(T_1 - T_0)\frac{1}{L} \tag{9.8}$$

where $\varkappa_{n,p}$ is the conductivity coefficient for an n-type and a p-type arm (leg), respectively; $S_{n,p}$ is the cross-sectional area of a

p-type and an n-type thermoelement arm, respectively; and L is the length of each arm.

Finally, the Joule heat may be written in the form

$$Q_J = I^2 L \left(\frac{\rho_n}{S_n} + \frac{\rho_p}{S_p}\right) \tag{9.9}$$

The mean amount of heat liberated at the hot end is

$$Q_J = \frac{1}{2} I^2 L \left(\frac{\rho_n}{S_n} + \frac{\rho_p}{S_p}\right) \tag{9.10}$$

The thermal equilibrium at the hot junction may be written as

$$Q_1 = (\alpha_{Tn} + \alpha_{Tp}) I T_1 + \frac{1}{L} (\varkappa_n S_n + \varkappa_p S_p)(T_1 - T_0) - \\ - \frac{1}{2} I^2 L \left(\frac{\rho_n}{S_n} + \frac{\rho_p}{S_p}\right) \tag{9.11}$$

In a similar way, the thermal equilibrium at the cold junction is

$$Q_2 = (\alpha_{Tn} + \alpha_{Tp}) I T_0 + \frac{1}{L} (\varkappa_n S_n + \varkappa_p S_p)(T_1 - T_0) + \\ + \frac{1}{2} I^2 L \left(\frac{\rho_n}{S_n} + \frac{\rho_p}{S_p}\right) \tag{9.12}$$

Substituting (9.11) and (9.12) into (9.5), we find the power taken by the load

$$P_l = I (\alpha_{Tn} + \alpha_{Tp})(T_1 - T_0) - I^2 L \left(\frac{\rho_n}{S_n} + \frac{\rho_p}{S_p}\right) \tag{9.13}$$

Since the total emf of the thermoelement is given by

$$E = (\alpha_{Tn} + \alpha_{Tp})(T_1 - T_0)$$

the voltage across the load is

$$V = E \frac{R_l}{R_l + R_L} \tag{9.14}$$

where the ohmic resistance of the thermoelement arms is

$$R_L = L \left(\frac{\rho_n}{S_n} + \frac{\rho_p}{S_p}\right) \tag{9.15}$$

The current in the thermoelement circuit can be written in the form

$$I = \frac{E}{R_l + R_L} = \frac{(\alpha_{Tn} + \alpha_{Tp})^2 (T_1 - T_0)}{R_l + R_L} \tag{9.16}$$

Substituting the expression for the current into (9.13) and introducing coefficient $m = R_l/R_L$, we obtain the following expression for the power in the load circuit:

$$P_l = \frac{(\alpha_{Tn} + \alpha_{Tp})^2 (T_1 - T_0)^2 m}{R_L (1 + m)} \tag{9.17}$$

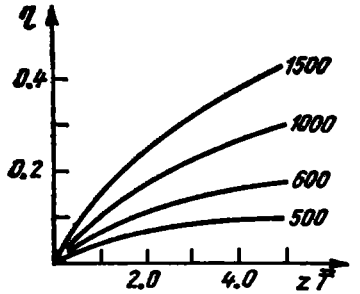


Fig. 9.4. Thermogenerator efficiency as a function of $z\overline{T}$ at various temperatures of the hot junction

It is easy to see that the power is maximum when $m = 1$. Hence,

$$P_{l\,\max} = \frac{(\alpha_{Tn}+\alpha_{Tp})^2 (T_1-T_0)^2}{4R_L} \tag{9.18}$$

The effectiveness of thermoelectric converters is characterized first of all by the efficiency η defined as the ratio of useful power P_l to the thermal power supplied to the thermoelement. In other words,

$$\eta = \frac{P_l}{Q_1} \tag{9.19}$$

Let us express Q_1 in terms of semiconductor parameters and temperature. For this purpose place into formula (9.11) expression (9.16) for the current:

$$Q_1 = \frac{(\alpha_{Tn}+\alpha_{Tp})^2 (T_1-T_0) I_1}{R_L (1+m)} + \\ + \frac{1}{L}(\varkappa_n S_n + \varkappa_p S_p)(T_1-T_0) - \frac{(\alpha_{Tn}+\alpha_{Tp})^2 (T_1-T_0)^2 m}{2R_L (1+m)^2} \tag{9.20}$$

The substitution of (9.17) and (9.20) into (9.19) and subsequent transformation give the following expression:

$$\eta = \frac{T_1-T_0}{T_1} \frac{\sqrt{1+z\overline{T}}-1}{\sqrt{1+z\overline{T}}+\frac{T_0}{T_1}} \tag{9.21}$$

where

$$\overline{T} = \frac{T_1+T_0}{2}$$

and

$$z = \frac{(\alpha_{Tn}+\alpha_{Tp})^2}{(\sqrt{\varkappa_n \rho_n}+\sqrt{\varkappa_p \rho_p})^2} \tag{9.22}$$

Figure 9.4 shows a plot of η as a function of $z\overline{T}$, calculated by formula (9.21). The value of z does not generally exceed 2×10^{-3} 1/°K and T_1 comes to about 600°K, with the result that $\eta \approx 10\%$.

Now we consider the requirements a semiconductor must meet when used in a thermoelement.

On putting in formula (9.22) $S_p = S_n = S$, $\alpha_n = \alpha_p = \alpha$, and $\rho_n = \rho_p = \frac{1}{\sigma}$, we obtain

$$z = \frac{\alpha_T^2 \sigma}{\varkappa} \qquad (9.23)$$

It is obvious from the above equation that the thermoelement efficiency rises with the electrical conductivity of a semiconductor since in this case the intrinsic resistance and losses due to Joule heat decrease, and conversely its efficiency drops off with an increase in the semiconductor heat conductivity as the flow of heat from the hot to the cold junction rises and so does the Q_1.

The condition would be ideal if the maximum electrical conductivity of a semiconductor went with its minimum heat conductivity. In actual fact, the values of $\varkappa$ and σ are closely interrelated and grow with increasing density of carriers. Expression (9.23) shows that neither metals (with low α_T) nor dielectrics (with low σ) are suited for thermogenerators. Fig. 9.5 illustrates the dependence of z on the density of free charge carriers in different substances. As is seen from the graphs, the values of z are maximum for semiconductors.

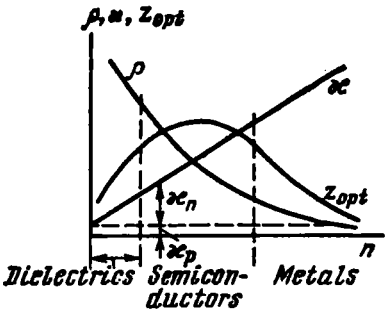


Fig. 9.5. Coefficient z_{opt}, resistivity ρ and heat conductivity $\varkappa$ *versus* carrier density

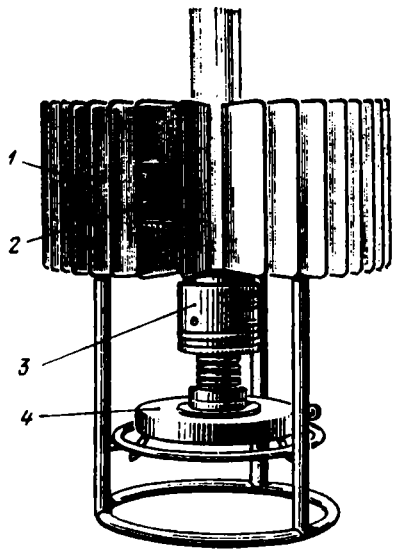


Fig. 9.6. External view of the ТГК-10 thermogenerator

1—fins; 2—terminals of thermoelements; 3—burner; 4—fuel container

The values of some parameters of semiconductors used in thermoelements are given in Table 9.1.

The power and emf of a thermogenerator can be increased by connecting thermoelements in parallel or in series. An example of such a device assembled of individual thermoelemets is the ТГК-10 thermogenerator turned out in the USSR. These generators with an

Table 9.1

Parameters of Semiconductors Used in Thermoelements

Semiconductor and its conductivity sign		α_T, μV/°C	σ, ohm⁻¹·cm⁻¹	ϰ, W/cm·°C	$z \times 10^3$, 1/°K
$CoSb_3$	(−)	200	400	400	0.4
ZnSb	(+)	200	350	200	0.7
PbS	(−)	160	600	220	0.9
PbSe	(−)	160	1100	240	1.35
PbTe	(−)	160	1500	280	1.5
Bi_2Te_3	(−)	170	1000	180	2.0
PbTe 2PbSe	(−)	160	900	120	1.8
$Bi_2Te_3-Bi_2(Se_3)$	(−)	170	1000	120	3.0
$Sb_2Te_3-Bi_2Te_3$	(+)	160	1500	140	2.8

efficiency of 3 to 3.5% (the external view is shown in Fig. 9.6) use the heat of an oil lamp. The hot ends of the generator are heated to 400-450°C, while the temperature of cold ends being cooled by convection of air reaches 90-100°C. These generators with an emf of about 1 V and load current up to 20 A can feed radio sets and small-power radio stations.

An important advantage of thermogenerators is their insensitivity to shorting.

A low efficiency of thermogenerators, to say nothing of their small life coming only to 4000-5000 h, does not allow them to compete with conventional electric energy sources, though intensive work being conducted in the field of thermoelectricity promises sharp increase in the efficiency of thermogenerators as soon as new semiconductors with high σ and small ϰ are made available.

Although the use of thermogenerators for commercial production of electric current is not yet economically advantageous, they can be effectively utilized even now, for example, in fire warning systems or in fire-hazardous difficult-to-control places or for the so-called cathode protection of underground metallic structures, in particular, oil- and gas-pipelines.

9.3. THERMOELECTRIC COOLERS

Thermoelectric coolers can be advantageously used in cooling small items.

Now we discuss how the current flowing through a thermoelement (Fig. 9.3) gives rise to a difference in temperature between the two junctions of the thermoelement.

The thermal balance between the thermoelement junctions results from the Peltier heat absorbed by the junctions and the Joule heat dissipated during the passage of current. Assuming that the *n*-type

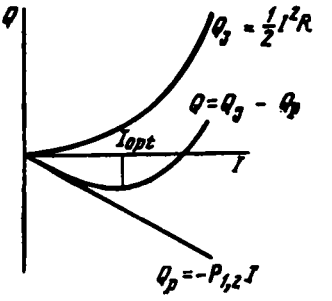


Fig. 9.7. Total heat absorbed at the thermoelement contact as a function of current through it

and p-type portions have identical resistances and that all the voltage drops off across the junctions, we may write that the resistance of each junction is 1/2 R. The Joule heat liberated at the junction will thus be equal to

$$Q_J = 1/2I^2R \tag{9.24}$$

where R is the total resistance of the thermoelement.

The sum of the Joule and the Peltier heat at the thermoelement junction is

$$Q_1 = -PI + 1/2I^2R \tag{9.25}$$

The dependence of Q_1 on the current through the thermoelement is given in Fig. 9.7. It is clear from the graph that the junction absorbs a maximum amount of heat at some optimum current I_{opt}. As the current grows the amount of the Joule heat begins to exceed that of the Peltier heat and the effect of cooling diminishes.

The optimum current I_{opt} may be determined from formula (9.25), for which purpose it must be analyzed for extreme values. As a result we obtain

$$I_{opt} = \frac{P}{R} \tag{9.26}$$

The optimum quantity of the heat drawn from junction *1* may be found by substituting (9.26) into (9.25):

$$Q_{1opt} = -\frac{P^2}{2R} \tag{9.27}$$

The temperature of the cold junction will continue to fall down during the heat withdrawal until the flow of the heat generated due to the Peltier effect becomes balanced out by the counterflow of heat which appears due to thermal conduction and which tends to equalize the temperatures at the cold and the hot junction.

The condition of thermal equilibrium is given by

$$Q_1 = \varkappa (T_1 - T_0) \tag{9.28}$$

By substituting (9.27) into (9.28), we obtain

$$(T_1 - T_0)_{max} = \frac{P^2}{2\varkappa R} \tag{9.29}$$

But as $P = \alpha_T T_0$ at the cold end and by formula (9.23) $z = \alpha_T^2/\varkappa R$, expression (9.29) takes the form

$$\Delta T_{max} = \frac{1}{2}\frac{\alpha_T^2}{\varkappa R} T_0 = \frac{1}{2} z T_0 \tag{9.30}$$

The expression thus shows that semiconductors to be used for thermoelectric cooling should also have the highest values of z.

Thermoelectric semiconductor coolers are used in quite a few industries. Thus, microcoolers intended for medical purposes (about one litre in volume) have proved quite economical and more convenient than compressor-driven coolers. The list of thermoelectric devices now in use is large enough; among these are: thermostats, ultrathermostats, microtomes, microscope stages, freezing traps for high-vacuum pumps, etc. We discuss here only two of them.

Microtomes are used in biological research to cut microscopic slices of tissue that must be first frozen to certain temperature.

In the conventional method, the required temperature is achieved with evaporation of carbonic acid. This method of cooling is inconvenient because it involves the use of carbonic acid compressed to 150 at. Besides, as this method of cooling practically excludes temperature control, only a trial and error approach can help catch the right moment at which the temperature of tissue has the desired value.

The use of a thermoelectrically cooled cabinet in the microtome has made it possible to dispense with unhandy and bulky high-pressure bottles and, above all, to control the cooling of tissue over a broad range of temperatures to a high accuracy by varying the current through the thermoelement.

Thermoelectric traps are used in vacuum technique. In rarefaction of gas to 10^{-7} mm Hg and below, use is made of oil-vapour pumps. To prevent oil vapour from penetrating into the space being evacuated, a special trap (gratings) cooled by liquid nitrogen is inserted between the pump and the space.

Cooling is provided by pumping liquid nitrogen in tubes that run through the gratings. If the vacuum installation runs for a long time care must be taken to timely fill in liquid nitrogen.

Thermoelectric traps now manufactured do not yet ensure so deep a vacuum as do liquid nitrogen coolers (grating temperature is not lower than —(35-50)°C at an ambient temperature of +20°C), but they are convenient in use and sufficiently economical as they consume merely 10W.

9.4. HALL-EFFECT DEVICES

When current flows through a sample placed in a transverse magnetic field, an emf is developed across the lateral faces of the sample at right angles to the current flow and to the magnetic field. We shall now discuss this phenomenon, called Hall effect.

The Lorentz force deflects conduction electrons in the magnetic field towards a lateral face of the sample. As a result, charges of opposite signs appear at the opposite lateral faces of the sample and thus an electromotive force is set up in proportion to the current and magnetic induction B (Fig. 9.8). In the absence of the load this emf may be written in the form

$$V_H = R\frac{IB}{d} \tag{9.31}$$

where R is the Hall constant; V_H is the Hall emf; and d is the sample width.

The expression for R can be obtained on the basis of the following reasoning. In the magnetic field, an electron is subjected to the action of the Lorentz force

$$f_H = q\bar{v}H$$

where $\bar{v}$ is the mean velocity of a charge carrier; and H is the field strength.

This force deflects the electrons to a lateral face, producing thereby an excess charge. The Lorentz force is cancelled by an electric field induced by the potential difference V_H:

$$f_E = q\frac{V_H}{d}$$

Under equilibrium conditions $f_E = f_H$, i.e.,

$$\bar{v}H = \frac{V_H}{d} \tag{9.32}$$

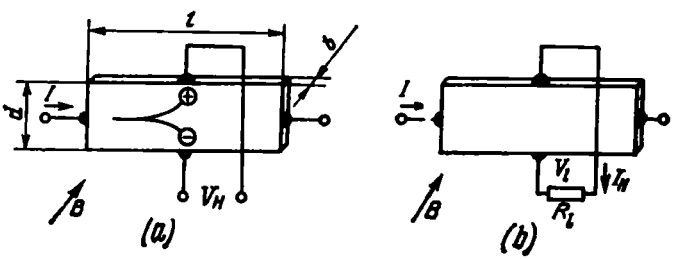


Fig. 9.8. Development of Hall emf in a semiconductor (a) no-load condition; (b) external load condition

The electric current flowing through the sample may be given by

$$I = dbnq\bar{v}$$

where n is the density of carriers; and db is the cross-sectional area of the sample.

Expressing $\bar{v}$ in terms of I and substituting it into (9.32) gives

$$V_H = \frac{1}{qn}\frac{IH}{d}$$

Comparing this expression with expression (9.31), we obtain

$$R = 1/qn \tag{9.33}$$

Bearing in mind that the mean carrier velocity depends on the mechanism of dissipation, expression (9.33) can be written in the form

$$R = A/qn$$

where $1 < A < 1.93$.

If we connect a load resistance R to the electrodes (Hall electrodes) attached to the lateral faces of the sample, it can be observed that a voltage V_l drops off across the load resistance. Therefore,

$$V_H = V_l + I_H R_H \tag{9.34}$$

where R_H is the sample resistance measured between the Hall electrodes in the magnetic field.

The current I_H in the load circuit is equal to V_l/R_l. Whence,

$$V_l = \frac{V_H}{1+\frac{R_H}{R_l}} \tag{9.35}$$

When the magnetic field is absent, $R_H = R_{H0}$ and expression (9.35) may thus be written as

$$V_l = \frac{V_H}{1+\left(\frac{R_{H0}}{R_l}\right)\left(\frac{R_H}{R_{H0}}\right)} \tag{9.36}$$

where R_H/R_{H0} is the relative variation in the sample resistance in the magnetic field.

It is clear from (9.36) that with the connection of the external load, the linearity between V_l and induction, that has been observed under the no-load condition, becomes disturbed now.

The Hall effect is widely used in determining the density and mobility of free carriers in semiconductors.

Since $R = A/qn$, then $R\sigma = A\mu$, and so, having measured conductivity σ and Hall constant R, we can find n and μ which are major parameters of a semiconductor.

The relation between V_H and B permits using the Hall effect in a sample with known parameters to measure the strength of a magnetic field.

The responsiveness of such a Hall generator is determined as a ratio between V_H and the value of B being measured:

$$\gamma_H = \frac{V_H}{B} \tag{9.37}$$

Itroducing here expression (9.31), we obtain

$$\gamma_H = \frac{RB}{Bd}\sigma E = \mu E \tag{9.38}$$

It is seen that the responsiveness of the Hall generator rises with an increase in the carrier mobility μ, for example, in generators made of InSb, in which mobility $\mu = 6000$ cm²/V·s, $\gamma_H = 2.10^6$ μV/Wb.

The Hall generator efficiency is determined as the ratio between the power taken by the external load and the input power. This quantity is called utilization factor. The input power delivered to the Hall generator is

$$P_{in} = I^2 R_H = I^2 \rho \frac{l}{bd} \tag{9.39}$$

When the Hall generator resistance R_H is equal to the load resistance R_l, the power taken by the oad is given by

$$P_H = \frac{V_H^2 ld}{4\rho b} \tag{9.40}$$

Substituting the expression for V_H from (9.31) into (9.40), we have

$$P_H = I^2 R^2 B^2 \frac{1}{4\rho bd} \tag{9.41}$$

The utilization factor $\eta = P_H/P_{in}$ can be found on dividing (9.41) by (9.39)

$$\eta = \frac{R^2 B^2}{4\rho^2} \tag{9.42}$$

For a semiconductor with one type of carriers, $R/\rho = A\mu$ and therefore

$$\eta = \left(\frac{A}{2}\right)^2 (\mu B)^2 \tag{9.43}$$

It follows from (9.37) and (9.43) that η is a maximum in semiconductors with a high mobility of carriers.

Table 9.2 gives some parameters of Hall generators made of different semiconductors.

Hall generators are successfully used for measuring large direct currents and powers and alternating currents and powers embracing

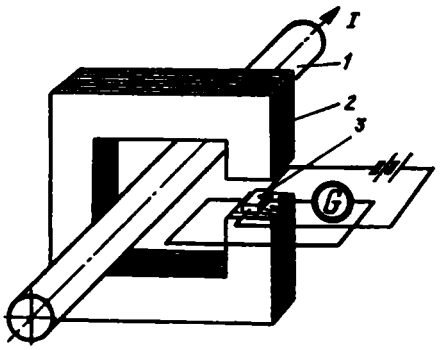


Fig. 9.9. Principle of operation of a Hall-effect noncontact ammeter

1—conductor carrying current to be measured; *2*—magnetic circuit; *3*—Hall generator

a broad frequency range up to high frequencies. If we place a conductor carrying current I to be measured inside a magnetic circuit with a gap in which a Hall generator is inserted (Fig. 9.9), then on determining magnetic induction $B \propto I$ from the Hall emf induced

Table 9.2

Parameters of Hall Generators Made up of Different Semiconductors

Semiconductor	Melting point, °K	ΔE, eV	σ, ohm^{-1}× ×cm^{-1}	R, cm^3/C	μ, cm^2/V·s	n, cm^{-3}	γ_H, μV× ×cm^2/Wb	η, %
Ge	1231	0.67	0.06	105	3800	10^{14}	540	0.05
InAs	1215	0.33	5	—900	30,000	10^{15}	500	3
InSb	796	0.18	100	—500	65,000	10^{16}	500	12.4
HgSe	963	0.23	500	—10	10,000	5×10^{17}	40	0.08
HgTe	943	0.45	1.5×10^3	—30	12,000	5×10^{18}	10	0.32

by current I in the magnetic circuit we thus find the value of I. In this way it is possible to carry out noncontact measurements of heavy currents. The Hall generator can serve as a wattmeter. The power taken by a load resistance through which current I_l is passed can be determined from magnetic induction $B \simeq AV_l$ produced in the gap of the magnetic structure, whose winding (*1* — *1* in Fig. 9.10) is fed with load voltage V_l, by measuring the voltage across the Hall electrodes of the generator placed in the gap and supplied with load current I_l:

$$V_H = AV_lI_l = AP$$

The wattmeters of this kind can operate at high frequencies up to a microwave range since the response time of the Hall generator is

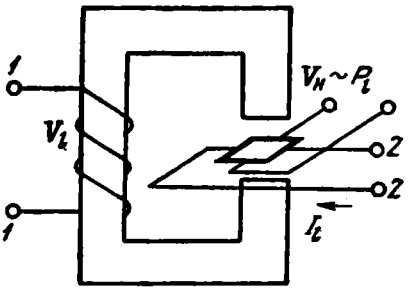


Fig. 9.10. Hall-effect wattmeter
1-1—load voltage; *2-2*—load current

very small, in the order of 10^{-14} s (it is determined by dielectric relaxation time $\tau = \varepsilon\varepsilon_0\rho$, i.e., the time of displacement of charges in the electric field).

The creation of semiconductors with high carrier mobilities, such as InSb, InAs, HgSe, and particularly of films of these semiconductors opens up new possibilities for application of Hall generators.

The ИМИ-3 induction measuring device manufactured in the USSR incorporates a semiconductor Hall generator. Hall generators are used to advantage in square-law detectors and frequency spectrum analyzers.

On setting up a magnetic field in the gap of the magnetic circuit by the current supplied to the Hall generator, we obtain the Hall voltage

$$V_H = \frac{RBI}{d} \approx RCI^2$$

where C is the factor of proportionality of B to I.

If the alternating current of frequency ω flows through the Hall generator, the Hall voltage is

$$V_H = \frac{RC}{d} I_m^2 \cos^2 \omega t = \frac{RC}{2d} I_m^2 + \frac{RC}{2d} I_m^2 \cos^2 \omega t$$

i.e., in this case, too, a direct current component of voltage is impressed on the electrodes, which is proportional to the square of the current amplitude.

Since the Hall voltage $V_H = (R/d)\, BI$, the Hall generator may be used as a linear detector in complex-shaped pulse spectrum analyzers. When the current to be analyzed is passed through a Hall generator and the magnetic field is set up by a sinewave voltage, then by varying the magnetic field frequency it is possible to measure the current amplitudes of all the harmonics entering into the spectrum of the pulse being analyzed because at any frequency the direct component of the voltage across the electrodes is equal to

$$V_{Hm} = \frac{R}{d} B_m I_m$$

Review Questions

1. Why is the thermal emf in metals lower than in semiconductors?
2. How does the sign of the thermal emf depend on the type of conductivity?
3. Why is the thermoelectric effect more pronounced at small currents?
4. What factors limit the value of thermal emf in thermogenerators?
5. What physical phenomena restrict the efficiency of thermogenerators?
6. Name the uses of thermoelectric current generators.
7. In which cases is thermoelectric cooling more economically advantageous than compressor-type cooling?
8. What determines the response of a Hall generator?
9. In which semiconductors does the Hall emf equal zero?

Chapter Ten

PHOTODEVICES

The class of photodevices embraces the apparatus that operate on the principle of interaction of light with semiconductor.

The absorption of light by semiconductors is associated with the transfer of radiant energy to the crystalline lattice. If the radiant energy $h\nu$ exceeds the work function $q\varphi$, i.e., is sufficient enough to be able to strip off an electron from the atom and emit it into a vacuum medium, then the effect of photoelectric emission sets in. This is the so-called extrinsic photoeffect which underlies the principle operation of vacuum photocells.

The intrinsic photoeffect involves the generation of electron-hole pairs in a solid body, which increase its electric conductivity. This phenomenon, also called photoconduction, is widely used in technology. The photoconductive effect is used in photoresistors, photodiodes, solar cells, phototriodes, etc.

10.1. PHOTOCONDUCTION

As is known, at any temperature a semiconductor has a certain amount of free electrons and holes in equilibrium, the density of the free electrons being dependent on the thermal generation of the carriers:

$$n_i = N_s e^{-\Delta E/2kT}$$

The absorption of light by a substance is associated with the transfer of radiant energy to a crystal lattice or to free electrons. The incoming radiant energy equal to or higher than the energy of the band gap width of a semiconductor ($h\nu \geqslant \Delta E$) causes the appearance of nonequilibrium electron-hole pairs in a solid body.

These pairs add to the semiconductor conductivity by the value

$$\Delta\sigma = q\,(\mu_n \Delta n + \mu_p \Delta p) \tag{10.1}$$

In semiconductors with a wide band gap, the density of Δn and Δp by far exceeds the equilibrium density of electrons and holes and therefore the photosensitivity of these semiconductors, defined as a ratio of the light conductivity to dark conductivity, is usually much higher than that in semiconductors with a narrow band gap width.

In metals whose intrinsic density of charge carriers is extremely high the photoconductive effect is nonexistent.

Since the mobilities of electrons and holes usually differ very much in one and the same semiconductor, it is possible to consider the carriers of only one sign in calculations. For the n-type semiconductor,

$$\Delta\sigma = q\mu_n \Delta n \tag{10.2}$$

The densities of nonequilibrium electrons and holes are related to incident light intensity L because they are proportional to the radiant energy absorbed per unit time per unit volume.

The energy absorbed by a unit area in a unit time is

$$-\frac{dL}{dx} = kL \tag{10.3}$$

where k is the absorption coefficient and x is the layer thickness.

The densities of nonequilibrium electrons and holes generated by the light in a unit time prove proportional to the absorbed radiant energy

$$\Delta p = \Delta n = \beta kL \ [\mathrm{cm^{-3} \cdot s^{-1}}] \tag{10.4}$$

The quantity β is called quantum yield because it determines on the average the number of electron-hole pairs generated by a quantum of radiant energy.

If the only consequence of illumination were the generation of electron-hole pairs, the conductivity of illuminated crystal would grow monotonically with time. But as the increasing density of electron-hole pairs is accompanied with a rise in their recombination, a steady value of the semiconductor photoconductivity corresponding to a given illumination intensity is set up in a certain time. At this steady value of the photoconductivity the rate of carrier generation and that of recombination are equal.

The steady carrier density, which has been established in the process of illumination and which determines the steady photoconductivity, must be proportional to the product of the densities Δp and Δn generated under illumination in a unit time by the lifetime of carriers up to their recombination τ, i.e., for electrons

$$\Delta n_{st} = \beta kL\tau_n \tag{10.5}$$

and for holes

$$\Delta p_{st} = \beta kL\tau_p \tag{10.6}$$

Quantities τ_p and τ_n depend on the process of recombination. It is obvious that the greater the number of holes in a semiconductor, the higher the probability of an electron running into a hole, and thus the less the time it stays in the conduction band until it recombines with a hole. As a result, the electron lifetime, or the time between the two collisions, is

$$\tau_n = \frac{1}{\gamma_n p} \tag{10.7}$$

and the hole lifetime is

$$\tau_p = \frac{1}{\gamma_p n} \tag{10.8}$$

where γ_n and γ_p are the recombination coefficients for electrons and holes, respectively.

The recombination coefficient depends on the thermal velocity $\bar{v}$ of an electron and capture cross section $q_{n,p}$:

$$\gamma_n = \bar{v}_n q_n$$

The cross section of electron capture by a trap centre is determined by the potential distribution near this centre.

The electron will be trapped by this centre if it approaches the centre for a distance at which the binding energy determined by electrostatic attraction will be equal to or more than the kT value. Then

$$\frac{q^2}{r\varepsilon\varepsilon_0} = kT, \quad r = \frac{q^2}{kT\varepsilon\varepsilon_0}$$

or at room temperature

$$q_{n,p} = \pi r^2 = \frac{10^{-10}}{(\varepsilon\varepsilon_0)^2} \text{ cm}^2$$

The value of $q_{n,p}$ may vary within a wide range from 10^{-1} to 10^{-25} cm^2 depending on the types of recombination centres. We now examine the types of recombination and their influence on the processes of photoconduction.

Linear recombination occurs when the concentration of recombination centres (holes, for example) is high and does not depend, to a first approximation, on illumination. In this case the electron lifetime according to (10.7) is

$$\tau_n = \frac{1}{\gamma_n p} = \text{const}$$

The value of the denominator in (10.7) determines the probability of recombination of one nonequilibrium electron. The recombination rate, or the number of carriers that recombine per unit volume per unit time, can be found on multiplying the recombination probability by the density of the carriers generated by radiation:

$$\Delta n q_n \bar{v} p = \frac{\Delta n}{\tau_n} \tag{10.9}$$

As a matter of fact, light generates both electrons and holes and we should have written the recombination rate in the form

$$\Delta n q_n \bar{v} (p + \Delta p)$$

But since it is the p-type semiconductor that is now being considered, in which p always much exceeds Δp, the value of Δp may be

dispensed with and the recombination intensity be by right expressed in the form of formula (10.9).

The analogous expression for the hole recombination rate takes the form

$$\Delta p q_p \bar{v} n = \frac{\Delta p}{\tau_p} \tag{10.10}$$

The recombination rate here proves proportional to the first degree of the density of carriers generated by light and therefore this type of recombination ($p \gg \Delta n$) is identified as linear recombination.

A change in the density of nonequilibrium carriers resulted from generation and recombination is described by the differential equation

$$\frac{d(\Delta n)}{dt} = \beta k L - \frac{\Delta n}{\tau_n} \tag{10.11}$$

The left-hand side of expression (10.11) refers to a change in the density of carriers due to generation and recombination. The first term of the right-hand side describes generation and the second term, recombination.

In solving equation (10.11) the initial conditions are so chosen that at the initial moment the excess carrier density is absent i.e., $\Delta n = 0$ when $t = 0$.

Dividing the variables and integrating, we get

$$\Delta n = \tau \beta k L (1 - e^{-t/\tau_n}) \tag{10.12}$$

When $t \to \infty$ the excess carrier density tends to an invariable value which determines a steady value of photoconductivity (10.5).

If a sample is not illuminated, the term in (10.11) which describes generation may be discarded and the expression can be written as

$$\frac{d(\Delta n)}{dt} = -\frac{\Delta n}{\tau_n} \tag{10.13}$$

For the initial condition $\Delta n = \Delta n_{st}$ and when $t = 0$, a fall in the photoconductivity can be found by integrating formula (10.13):

$$\Delta n = \Delta n_{st} e^{-t/\tau} \tag{10.14}$$

Relaxation curves describing the fall and rise of photoconductivity in case of illumination by a square pulse are shown in Fig. 10.1.

By relaxation is meant the transient time required for the photoconductivity to reach a steady value after the light has been switched on or the time it takes the conductivity to fall to a steady value of the dark conductivity after turning off the light.

Figure 10.1 shows that the rise time and the fall time of photoconductivity are equal.

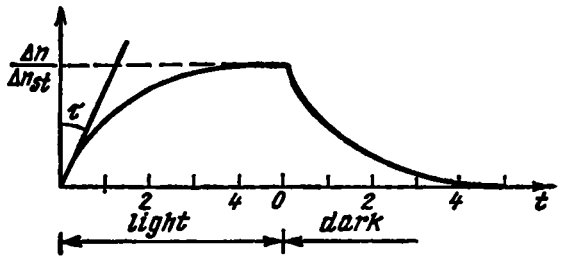


Fig. 10.1. Curves of photocurrent rise and fall in case of linear recombination under pulse illumination of a sample

The relaxation constant for linear recombination coincides with the lifetime of carriers and consequently, by (10.14), it becomes possible to determine experimentally $\tau_{n,\,p}$ from the plot of function $\ln\left(\frac{\Delta n}{\Delta n_{st}}\right) = f(t)$.

Quadratic recombination is observed when the density of recombination centres (holes, for example) is low. The recombination of nonequilibrium electrons is governed in this case by the number of holes excited during illumination. Since the densities of nonequilibrium electrons and holes are equal, the rate of recombination can be determined by substituting $\Delta p = \Delta n$ into (10.9) for p. In other words, the recombination rate can be written in the form $(\Delta n)^2\gamma$.

The recombination rate here is proportional to the square of the carrier density and therefore this recombination is called quadratic.

Now we examine relaxation processes during quadratic recombination. Time variations in the carrier density during illumination and recombination may be written as

$$\frac{d(\Delta n)}{dt} = \beta kL - \gamma_n (\Delta n)^2 \qquad (10.15)$$

As in the case of linear recombination, the initial conditions for the integration of (10.15) have the form

$$\Delta n = 0 \text{ when } t = 0$$

Integrating (10.15) gives

$$\Delta n = \sqrt{\frac{\beta kL}{\gamma_n}} \tanh t\sqrt{\beta kL\gamma_n} \qquad (10.16)$$

The equation describing the current fall after switching off the light takes the form

$$\frac{d(\Delta n)}{dt} = -\gamma_n (\Delta n)^2 \qquad (10.17)$$

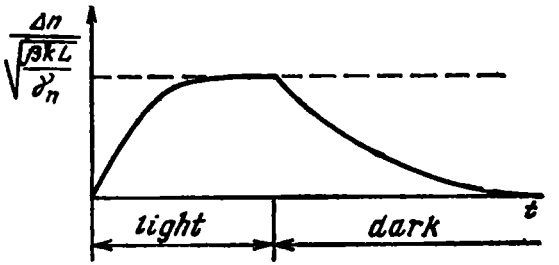


Fig. 10.2. Curves of photocurrent rise and fall in case of quadratic recombination during pulse illumination of a sample

Integrating (10.17) for the initial conditions of $t = 0$ and $\Delta n = \sqrt{\frac{\beta k L}{\gamma_n}}$, we get

$$\Delta n = \sqrt{\frac{\beta k L}{\gamma_n}} \frac{1}{t\sqrt{\beta k L \gamma_n} + 1} \tag{10.18}$$

The curves of the rise and fall of photoconductivity plotted by formulas (10.16) and (10.18) are given in Fig. 10.2.

As is seen from the graph, the rise time of the photoconductivity up to its stationary value and the fall time of the photoconductivity down to a steady dark conductivity value are not equal (as distinct from linear recombination).

The number of recombination centres and the lifetime of carriers vary uninterruptedly in the process of quadratic recombination. In contrast to linear recombination, an instantaneous lifetime of carriers has a meaning here

$$\tau_t = \frac{1}{\gamma_p \Delta p(t)}$$

Needless to say that the above reasoning is very much simplified. In actual fact, the magnitude and kinetics of photoconductivity are heavily dependent on the traps resulted from the impurities and defects of the crystal lattice.

Carrier traps present in a semiconductor can reduce the photoconductivity hundreds of times and considerably increase its relaxation time. Traps are conducive to the so-called infrared decay of photoconductivity. If the band gap has shallow traps located near the bottom of the conduction band, they may be filled in at room temperature and will thus exert no influence on the photocurrent value.

It should be noted that the energy of the emitted light is related to its wavelength by

$$E\,[\text{eV}] = \frac{1.23}{\lambda\,[\mu\text{m}]}$$

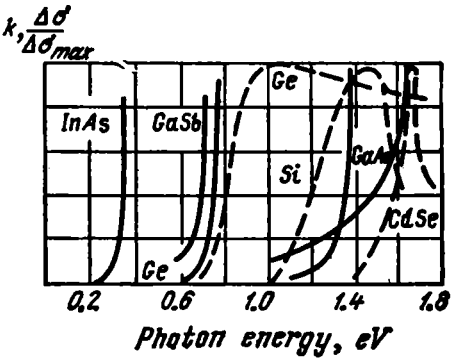


Fig. 10.3. Spectral distribution of the absorption coefficient (solid lines) and of photoconductivity (dotted lines) in certain semiconductors

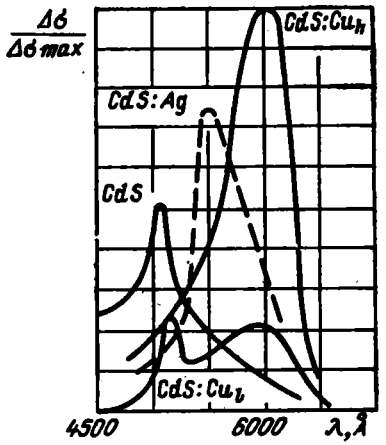


Fig. 10.4. Spectral distribution of impurity photoconductivity in CdS (subscripts *h* and *l* denote samples with high and low concentrations of copper, respectively)

For example, cadmium sulphide with $\Delta E \simeq 2.46$ eV displays photoconduction properties when exposed to light with a wavelength of less than 0.5 μm. If in addition to this irradiation the semiconductor is illuminated by infrared light whose quantum energy is sufficient to free electrons from traps though not high enough to transfer electrons from the valence band to the conduction band, then passage of the electrons excited by the light to the vacant levels of the traps will cause a sharp decrease in the photocurrent.

Absorption spectra and spectral distribution of photoconductivity in some semiconductors are plotted in Fig. 10.3. The Figure shows that the plot of photoconductivity as a function of radiant energy has a maximum at the energy level which is slightly in excess of the band gap energy level and is conducive to a sharp rise in the absorption coefficient.

Up till now the discussion has been devoted to intrinsic photoconductivity, in which case the maximum photosensitivity corresponds to the radiant energy that exceeds the energy of the semiconductor band gap.

Impurities present in a semiconductor can sharply change the pattern since light absorption by impurity centres will give rise to their ionization and on the curves of spectral distribution of photoconductivity there will appear maxima with energies corresponding to the impurity activation energy (Fig. 10.4).

10.2. PHOTORESISTORS

It has been mentioned in the previous Section that the lower the initial conductivity of a semiconductor, the higher its photosensitivity specified as a ratio between the light and dark conductivities.

For this reason photoresistors are prepared from semiconductors with a low density of impurities.

The spectral sensitivity range of a photoresistor is determined by the band gap width and impurity levels in a semiconductor. Table 10.1 lists wavelengths λ_{max} corresponding to the maxima of spectral sensitivity, and also band gap widths of semiconductors used in the manufacture of photoresistors.

Table 10.1

Light Wavelengths Corresponding to Maxima of Spectral Sensitivity in Semiconductors Used for Photoresistors

Material	ΔE, eV	λ_{max}, μm	Type of photoresistor	Material	ΔE, eV	λ_{max}, μm	Type of photoresistor
PbS	0.4	2.5	ФСА	CdS	2.4	0.5	ФСК
Bi_2S_3	1.3	0.95	ФСБ	CdSe	1.7	0.7	ФСД

The maxima given in the Table correspond to the photosensitivity displayed in the region of intrinsic conduction. These maxima may be shifted and the sensitivity range expanded by introducing impurities in semiconductors.

Photoresistors are generally prepared by sintering a powder of photosensitive material or by depositing a semiconductor film on an insulating substrate (by vacuum evaporation or chemical deposition).

Photoresistors are sometimes made of monocrystal wafers and the type notation of these photoresistors includes letter M. Thus, the photoresistor made from a cadmium sulphide monocrystal is designated ФСК-М1. To provide for easy use and protection against exposure, the semiconductor wafer with the electrodes deposited

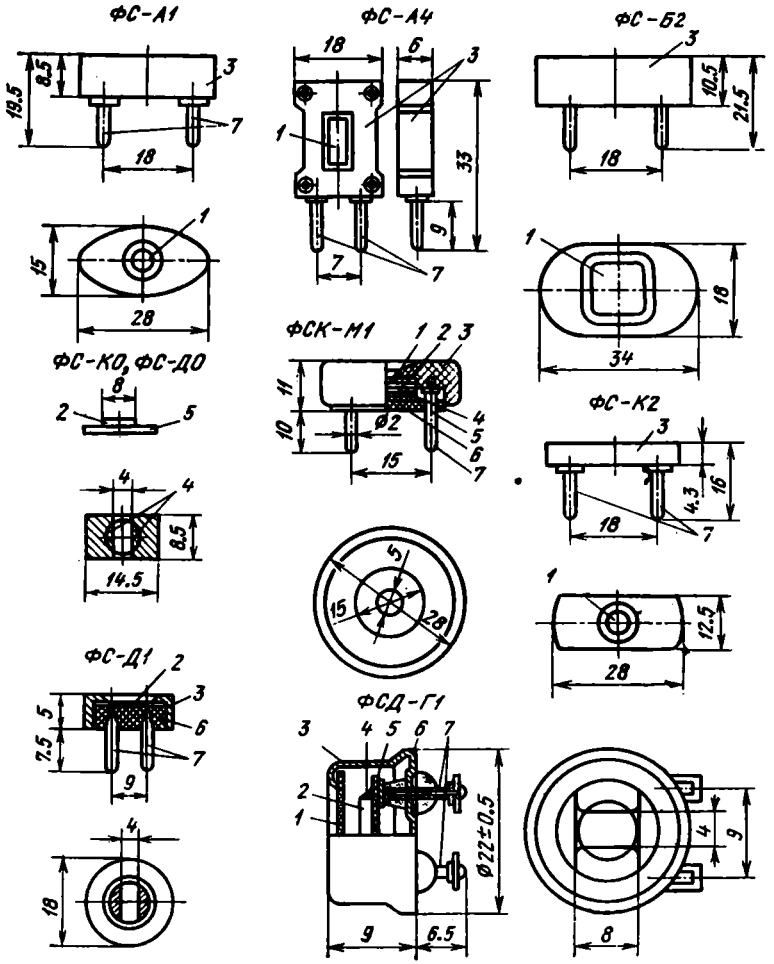


Fig. 10.5. Design configurations of photoresistors

1—window in the case; *2*—photosensitive element; *3*—case; *4*—contact attached to the semiconductor slice; *5*—dielectric substrate; *6*—case cover; *7*—terminals

thereon is placed into a plastic case having a window closed by glass, mica or varnish, which are transparent in the range of spectral sensitivity of the semiconductor. Fig. 10.5 shows a number of the design configurations of photoresistors. Some photoresistors are hermetically sealed and the type notation of such photoresistors includes letter Г. For example, the encapsulated cadmium sulphide photoresistors are marked ФСК-Г.

The photoresistor performance is usually characterized by the value of photocurrent (light current) I_l flowing through the photoresistor at a given voltage and light flux Φ.

The light flux (in lumens) incident on a photoresistor with effective area S (cm^2) and illumination L (in luxes) is determined from the formula

$$\Phi = L \cdot S$$

This photocurrent represents a difference between the dark and light currents. Besides, the dark current I_d at a given voltage and the multiplicity of changes in the resistance are also specified:

$$\frac{\Delta R}{R} = \frac{R_d - R_l}{R_l}$$

where R_l is the light resistance and R_d is the dark resistance of the photoresistor.

The value of R_d varies with different types of photoresistors from 10^7 to 10^{11} ohms. R_d is particularly high (10^{11} ohms) in cadmium sulphide monocrystal photoresistors.

Depending on their design, the photoresistors are differentiated by the maximum values of permissible power P_{max} and impressed voltage V_{max}, the rise of which causes irreversible changes in the devices due to heating or punch-through.

The integral specific sensitivity of a photoresistor is determined as a ratio of the photocurrent to the product of the incident luminous flux and applied voltage:

$$K_0 = \frac{I_l}{\Phi V}$$

where Φ is the luminous flux produced by an incandescent lamp with a tungsten filament heated to 2840°C. Such "standardization" of the glow temperature permits of comparing the sensitivity of differing photoresistors by the light flux with an invariable spectrum.

The sensitivity K_0 rises with lowering temperature as this entails a decrease in the density of equilibrium electron-hole pairs and thus a fall in the dark conductivity and recombination probability.

The spectral distribution of sensitivity, being a function of the band gap width and impurity density in a semiconductor, is specified for each photoresistor.

The light pulse frequency at which the photocurrent amplitude does not yet depend on the frequency is determined by the lifetime of carriers (see Sec. 10.1). If the duration of the light pulse, t_l, and the interval between the light pulses, t_d, do not exceed τ, the photoconductivity diminishes because during the light pulse the photocurrent has no time to reach a steady value while the interval between the light pulses is not long enough for the photocurrent to drop off to an equilibrium dark value.

Photoresistors of certain types have appreciably low lifetimes of carriers (10^{-5} s for the ФСА type).

One should point out that the time of relaxation and the sensitivity of photoresistors are interdependent. The higher the relaxation time, the larger the value of K_0. This is obvious as both K_0 and time of response rise with increasing lifetime of carriers.

Table 10.2

Parameters of Photoresistors

Parameter	ФСА-1	ФСВ-2	ФСК-М1 ФСК-С2	ФСК-1	ФСК-Г2	ФСД-0 ФСД-1
R_d, ohms	10^4-10^6	10^5-10^7	10^8-10^{11}	10^7	1.6×10^6	10^5 to 10^8
P_{max}, W	0.01	—	—	0.1	0.2	0.05
V_{max}, V	15	50	130	400	250	300
λ_{max}, μm	2.1	0.7	0.5	0.6	0.5	0.75
K_0, mA/V·lm (under small illumination)	0.5	1	50	6	2.5	25
τ, ms (with small light fluxes)	4×10^{-2}	1	100	25	30	40
Photosensitivity area S, cm²	0.3	1.2	0.1	0.3	0.3	0.3
Multiplicity ratio of variation in resistance, R_d/R_l, at 200 lx	12	—	100	100	80	150
Dark current I_d, μA, at 50 V	—	—	—	15	no more than 15	no more than 10

Some of the parameters of photoresistors produced in the USSR are given in Table 10.2. Photoresistors are used in direct and alternating current circuits of measuring devices and apparatus of photoelectric automatic systems.

10.3. PHOTODIODES AND PHOTOCELLS

Photosensitive *p-n* junction devices, such as photodiodes, photocells and phototransistors are extensively used in various branches of technology. For example, photodiodes are applied in various automatic control circuits and photocells (solar cells) are used in space engineering, photometry, etc.

Photodiodes

It has been shown in Chapter Two that the *p-n* junction reverse current is governed by the density of minority carriers in the *p-n* junction region. On illumination of the *p-n* junction by the light

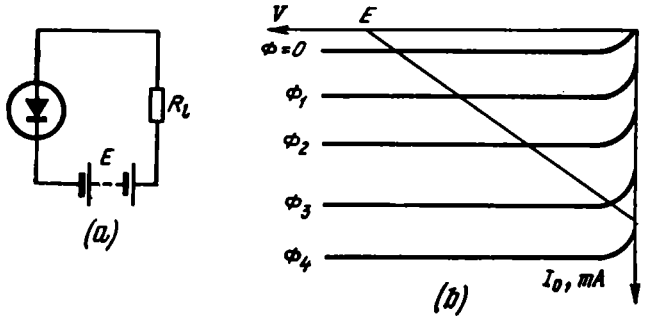


Fig. 10.6. Connection diagram (*a*) and current-voltage characteristics of a photodiode at different levels of illumination intensity (*b*)

with a sufficient energy of photons ($h\nu \geqslant \Delta E$), electron-hole pairs begin to appear in the immediate vicinity of the *p-n* junction and the density of minority and majority carriers grows near the junction.

The current in such a reverse-biased junction rises steeply and the current-voltage characteristic takes the form shown in Fig. 10.6*b*.

The light source injects electrons and holes in the semiconductor, which are collected at the *p-n* junction. The saturation current I_0 depends on the value of light flux Φ.

Let us draw a load line on the current-voltage characteristic of the photodiode. In the absence of illumination the circuit carries a small current which is determined by the reverse resistance of the diode (about 1-2 μA for silicon photodiodes and 15-20 μA for germanium photodiodes). In this case all the voltage drops off across the photodiode.

On illumination of the photodiode, the current through the *p-n* junction rises to the value $I_l > I_d$. The current I_l is produced here by the light-excited carriers. Almost all the voltage falls across the load resistance. The voltage drop across the photodiode may be disregarded with a sufficiently large load resistance R_l.

The current-voltage characteristic of a reverse-biased junction may be written in the form

$$I = I_0 (e^{qV/kT} - 1) - I_l \qquad (10.19)$$

As in the case of photoresistors, the integral photosensitivity of photodiodes is characterized by the photosensitivity coefficient

$$K_{ph} = \frac{dI_l}{d\Phi} \text{ mA/lm}$$

The integral photosensitivity of a photodiode does not depend on voltage as the photodiode operates in the conditions of current satu-

ration (see Fig. 10.6). As a rule, until the illumination reaches very high values the photocurrent in the photodiode generally rises linearly with increasing light flux.

Photodiodes are also characterized by the threshold sensitivity, i.e., the minimum illumination under the action of which one can observe variations in the external circuit current, recognizable against the background of the internal noise. Some of the parameters of photodiodes are presented in Table 10.3.

Table 10.3

Parameters of Photodiodes

Parameter	Photodiodes			
	ФД 1	ФД 2	ФД 3	ФД К
Working voltage, V	15	30	10	20
Maximum working voltage, V	20	50	—	—
Dark current, μA	30	25	15	3
Integral sensitivity K, mA/lm	20	10-20	20	3
Photosensitivity area, cm^2	0.05	0.01	0.02	0.02

A high efficiency of a photodiode is achieved when most of the radiant energy is absorbed near the *p-n* junction.

If the electron-hole pairs are generated at a distance from the *p-n* junction that considerably exceeds the carrier diffusion length, the carriers will have time to recombine before reaching the *p-n* junction and there will be no perceptible rise in the current through the diode.

This is why the photodiode with the light beam directed parallel to the junction plane has not found application as the carriers are generated in this case in the near-surface layer and the junction area is not fully utilized.

These shortcomings are obviated if the light that impinges on the junction is directed at right angles to its plane. The other condition for the full utilization of the junction area is a small depth (smaller than $L_{p,n}$) of penetration of the *p-n* junction into a semiconductor body. Fig. 10.7 shows the structure of ФД-1 and ФД-3 photodiodes manufactured in the USSR.

The photodiode header (crystal holder) has a window closed by a lens focusing the light onto a germanium wafer. A tin ring placed underneath the *n*-region serves as its lower contact. The case may be made of metal, as in the ФД-1 and ФД-3 types, or of plastic, as in the ФД-2. The transparent plastic case is covered with light-tight varnish, leaving intact a transparent window opposite the photosensitive surface.

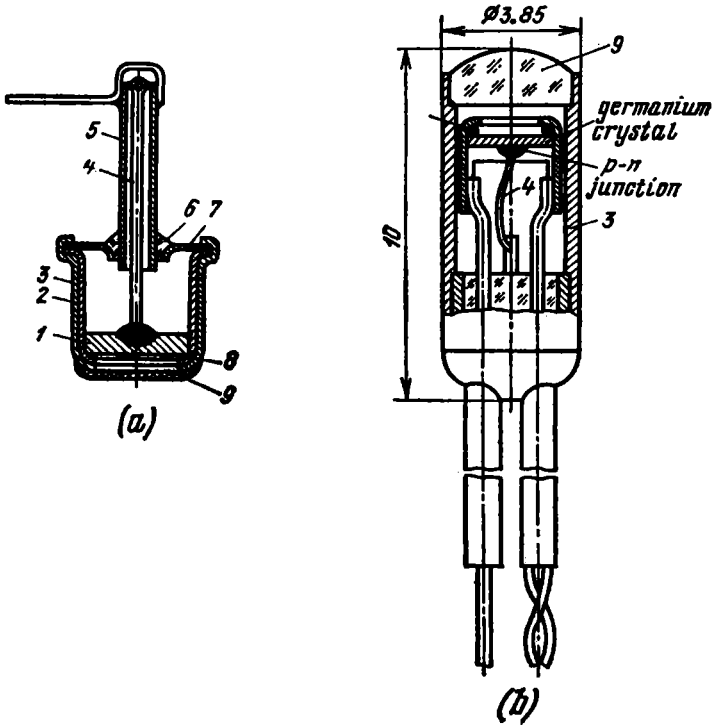


Fig. 10.7. Photodiode designs, type ФД-3 (*a*) and ФД-1 (*b*)
1—germanium slice; *2*—crystal holder; *3*—kovar can; *4*—lead; *5*—kovar tube; *6*—glass insulator; *7*—case stem; *8*—tin ring; *9*—glass lens

As with photoresistors, the spectral sensitivity of photodiodes is determined by the band gap width of a semiconductor. Thus, for germanium, this is a visible and a short-wave region of the infrared spectrum and for silicon, a visible light spectrum.

The frequency response of a photodiode depends not on the carrier lifetime but rather on the time it takes the carriers to diffuse from the surface into the *p-n* junction region:

$$\tau = \frac{x^2}{D_n}$$

where x is the *p-n* junction depth.

The working frequency of diffused photodiodes with $x = 3$ to 5 μm reaches 10^7 Hz and even more. Photodiodes are extensively used in engineering. Thus, the ФД-3 type germanium photodiodes are used to advantage as near infrared radiation detectors. Compared to photoresistors, photodiodes are more widely used in automatic circuits as photoelectric converters owing to the fact that under

illumination the intrinsic resistance of photodiodes becomes sufficiently small. When series-connected to the load resistance R_l in such a circuit, the photodiode is able to control the load current over a wide range.

Photocells

On illuminating a *p-n* junction, the light-generated electron-hole pairs get divided by the contact field of the junction—the holes are pulled over to the *p*-region and the electrons, to the *n*-region, with the result that an additional space charge appears on both sides of the *p-n* junction and thus an emf develops between the *p*- and *n*-regions in the absence of an external voltage.

Certainly, the carrier storage does not proceed without end as it is limited by the diffusion of majority carriers across the *p-n* junction (see Ch. Two).

If we connect a load resistance to the terminals of such a photocell, a current will flow through the resistance. The current-voltage characteristics of a photocell are shown in Fig. 10.8. The points of intersection of each curve with the current axis correspond to short-circuit conditions ($R_l = 0$) and with the voltage axis, to open-circuit conditions ($R_l = \infty$).

The open-circuit voltage of the photocell can be easily found from (10.19) by setting $I = 0$

$$V_{oc} = \frac{kT}{q} \ln \left(\frac{I_l}{I_0} + 1 \right) \tag{10.20}$$

This voltage cannot exceed the value numerically equal to the value of the band gap width of a semiconductor. The V_{oc} of photocells comes to approximately 1 V.

In commercial production are silicon and selenium cells which enjoy wide use. Solar batteries composed of silicon cells with large photosensitive areas are employed as power sources of space equip-

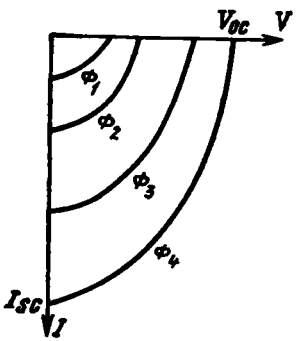


Fig. 10.8. Current-voltage characteristics of a photocell at different levels of illumination intensity

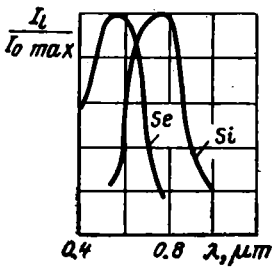


Fig. 10.9. Spectral distribution of photosensitivity for silicon and selenium photocells

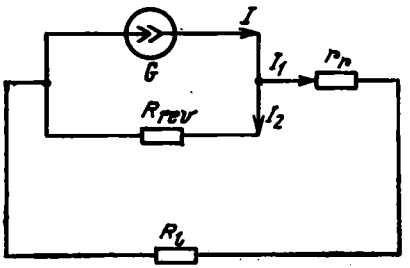


Fig. 10.10. Photocell equivalent circuit

r_p —resistance of a semiconductor crystal; G—current generator; R_l —load resistance

ment, and polycrystalline selenium cells are applied in photometry devices, exposure meters, etc.

As for their design and manufacture, the photocells differ from the diodes and rectifiers made of the same materials only in the method of fabrication of the contact attached to a photosensitive layer. In a silicon cell, this can be a metallic ring if the photosensitive region is low-resistive, or a continuous translucent layer of metal which is deposited on a high-resistivity semiconductor. In a selenium cell with a high-resistivity photosensitive region, the contact is made in the form of a translucent metallic film (commonly gold).

The spectral response (sensitivity characteristic) of photocells coincides with that of photodiodes, and for silicon cells its maximum corresponds to wavelengths of 0.7 to 0.8 μm, which is close to the maximum of the spectral distribution of the solar light energy. In selenium cells the spectral maximum embraces a wavelength range of 0.5 to 0.7 μm (Fig. 10.9). An interesting feature of selenium photocells is the fact that their spectral response characteristic is close to that of the man's eye, which permits using them in photometry.

At small light intensities the photocell current varies linearly with the light flux, but when the light intensity is greatly increased the linearity is disturbed due to the dependence of the *p-n* junction resistance on illumination intensity.

Let us consider a photocell equivalent circuit (Fig. 10.10). Current I generated by light (corresponding to a current generator G in the equivalent circuit) is proportional to the light flux, $I = A\Phi$.

At the branch point of the current

$$\frac{I_2}{I_1}=\frac{R_l+r_r}{R_{rev}}$$

but since $I = I_1 + I_2$, the current in the external circuit is

$$I_1=\frac{I}{1+\frac{R_l+r_r}{R_{rev}}}=\frac{A\Phi}{1+\frac{R_l+r_r}{R_{rev}}}$$

As $R_l = 0$ in the short circuit conditions, then

$$I_1=\frac{A\Phi}{1+r_r/R_{rev}} \tag{10.21}$$

At small levels of illumination intensity, $R_{rev} > r_r$ and $I_1 = A\Phi$. At higher illumination levels, R_{rev} begins to diminish and the dependence of I_1 on Φ is no longer linear.

The effectiveness of the photocell is described by its efficiency which is the percentage ratio between the maximum energy given up to the external circuit and the total radiant energy that falls on the photocell. The efficiency of silicon cells is 10-15% and that of selenium cells, 0.2%. The low values of the efficiency are due to the radiant energy losses that result in decreased generation of electron-hole pairs. These losses may be divided into two groups:

1. Radiant energy loss by reflection, inactive light absorption (without generation of carriers) and light absorption at a large distance from the *p-n* junction.

2. Energy loss due to recombination of electron-hole pairs outside the *p-n* junction, presence of series resistance r_r and parasitic loss. For silicon cells, the first-group losses come to about 35% and the second-group losses, 55%.

The efficiency of photocells diminishes with increasing temperature. The temperature coefficient of silicon cells is 5×10^3 1/°C.

10.4. PHOTOTRANSISTORS

Photodiodes and photoresistors do not amplify the photocurrent they generate. Unlike these types of photodevices, the phototransistors are active converters of radiant energy.

The phototransistor is a common junction triode connected into a circuit as is shown in Fig. 10.11*a*.

The collector dark current I_{c0} in such a triode is higher than that in a diode

$$I'_{c0}=\frac{\alpha I_{c0}}{1-\alpha}$$

On illuminating the base region, the generated pairs of electrons and holes become separated near the collector junction. The holes are drawn to the collector and the electrons remain in the base.

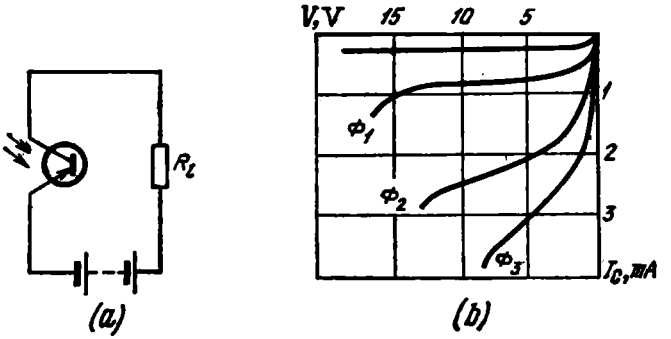


Fig. 10.11. Connection d agram (*a*) and a set of collector current-voltage curves (*b*) for a phototransistor under different illumination intensities

When the base operates in the zero-current mode the excess charge of electrons in the base is not allowed to decrease due to increasing base current.

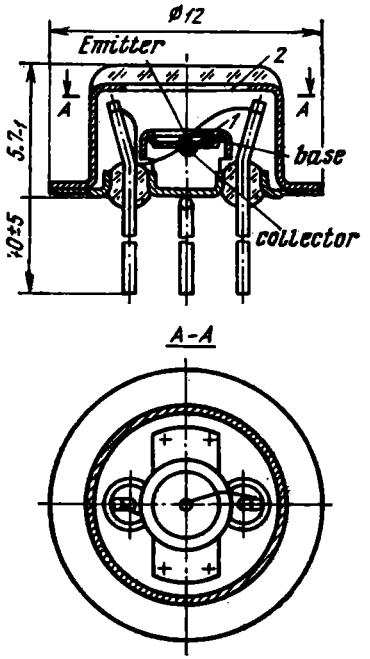


Fig. 10.12. ФТ-1 phototransistor structure

1—semiconductor crystal; 2—window in the transistor case

With an increase in the number of electrons in the base, the barrier across the emitter junction is lowered by the field of this space charge. An additional number of holes are then injected from the emitter into the base, most of which reach the collector and increase its current (Fig. 10.11*b*).

Thus, the light-generated space charge present in the base causes amplification of the photocurrent in the collector *p-n* junction.

The variation in the current flowing in the external circuit is

$$I_{ph} = \frac{I_l \alpha}{1-\alpha} = \beta I_l$$

The phototransistor sensitivity is thus governed by the base region photosensitivity and the transistor current-transfer ratio β. Owing to the current amplification, the integral sensitivity can be brought up to hundreds of milliamperes per lumen.

In germanium phototriodes of the ФТ-1 type the base does have a terminal after all, which is used to produce a bias required for obtaining a linear dependence and offsetting temperature variations in the current. The design of this triode is shown in Fig. 10.12. Its sensitivity $K_{ph} = 0.17$ to 0.5 A/lm and the lumen equivalent of noise, $F_{lm} = (5 - 100) \times 10^{-8}$ lm at $V_n = 5$ mV

10.5. SEMICONDUCTOR LIGHT SOURCES

Extensive use has been recently made of a luminescence phenomenon in various branches of science and engineering. Luminescence is observed in gaseous, liquid and solid substances and represents a substance-emitted electromagnetic radiation with a wavelength corresponding to the light or near-the-light wavelength range, and not associated with heat. The luminescent radiation is thus a "cold" radiation excessive with respect to the equilibrium thermal radiation given off by the substance.

The luminescence effect is based on the ability of electrons in a solid body to be excited or imparted an additional energy under the action of a certain external energy source so that the transition of electrons to higher empty energy levels results. This state is not stable (if it is not associated with heating of the body), and upon termination of the excitation the electrons return to the lower unoccupied levels. A part of the electrons usually expend their energy directly on heating the body within which this process takes place. The other portion of the electrons can give off the energy in the form of quanta of electromagnetic radiation (photons). It is the flow of photons that is observed as luminescent radiation.

Luminescence may be classified by the methods of electron excitation. Most common are photoluminescence (excitation by light), electroluminescence (excitation by electric field) and cathodoluminescence (excitation by an electron beam).

These types of luminescence are widely used in practice, for example, in various electro- and photo-luminescent devices intended to produce images (signal panels, sign indicators and luminescent scales), image translators and image amplifiers (including snooperscopes, i.e., night viewing devices, and scintillation counters), television picture tubes (kinescopes), discharge lamps, semiconductor light-emitting diodes and, lastly, lasers of various types (semiconductor lasers included).

The luminescence properties exhibited by semiconductors are due to the transition of electrons from the valence band or from the impurity levels to the conduction band. These excess nonequilibrium electrons perform reverse quantum transitions to the valence band (or to the impurity levels) and recombine with holes. The quantum transitions are divided into radiative and nonradiative ones depend-

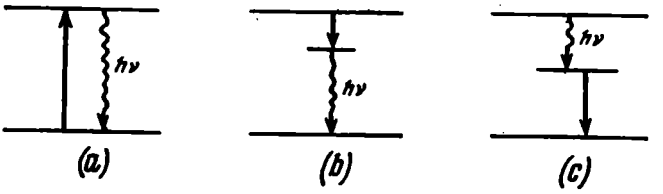


Fig. 10.13. Scheme of possible radiative transitions in a semiconductor
(a) direct; (b) via impurity centre; (c) lattice defect

ing on whether these transitions are attended or not with emission of photons. The emitted radiation is called recombination emission. In some of the semiconductors radiative quantum transitions take place directly from the conduction band to the valence band; in others, this kind of transitions is impossible and the transitions are executed only through the intermediary of one or two impurity levels or those of lattice defects located in the band gap. The scheme of possible quantum transitions in the luminescence process is shown in Fig. 10.13. The nonradiative transitions (in which case the liberated energy is expended on heating the crystal lattice) occur either through impurity levels (or defect levels) or as a result of the transfer of the energy released in recombination to conduction band electrons which in turn expend it on heating the crystal lattice.

This is why one of the main technological problems in the preparation of effective devices operating on the principle of recombination radiation is the provision of favourable conditions for the initiation of radiative quantum transitions (precision alloying with chosen impurities) and removal of unwanted impurities and defects which facilitate the occurrence of nonradiative transitions.

This problem is as a rule difficult to solve for semiconductors with broader band gaps due to their high melting points. Generation of an intensive recombination luminescence becomes, therefore, a complex problem with a decrease in the light wavelength.

Semiconductor emitters which depend for their operation on the effect of electroluminescence are the best known and most extensively used devices. Among these are phosphor-based electroluminescent cells. Phosphors are prepared in the form of fine-crystalline powders suspended in a dielectric medium (epoxy resins, glasses) or in the form of thin, homogeneous, sublimated semicrystalline films (sublimate-phosphors). Phosphors are generally made from type $A^{II}B^{VI}$ compositions and primarily from zinc sulphide with different admixtures (activators). These phosphors emit light in the visible spectrum within the range of wavelengths from 4500 to 6000 Å depending on the admixtures introduced into them

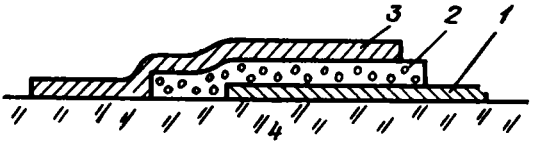


Fig. 10.14. Structure of the electroluminescent film cell of an indicator

1—lower electrode; *2*—film of a semiconductor light emitter; *3*—translucent electrode; *4*—substrate

The structure of an electroluminescent film cell is given in Fig. 10.14. Thin metallic layer *1* serving as an electrode is deposited on glass or ceramic substrate *4*. Phosphor layer *2* (as ZnS : Mg) is obtained by vacuum evaporation, on which is sputtered a translucent film used as second electrode *3*.

Powdered luminophors emit light only on applying an alternating voltage of ≈150-200 V with a frequency of 0.4-20 kHz to the electrodes of a cell. Such luminophors offer little promise. Sublimate-phosphors fluoresce at smaller voltages—from 2-2.5 to 30-60 V and in a number of cases they display luminescence both at alternating and direct voltages. But the life of sublimate-phosphors is so far very short (≈200 h) and the reliability is poor.

More promising are semiconductor emitters whose operation is based on the effect of injected electroluminescence. These are light-emitting diodes or semiconductor lamps in which radiation originates through the recombination of electrons injected in the *p*-region by the *p-n* junction biased in the forward direction. These injected electrons recombine in the narrow layer adjacent to the junction, the layer width being of the order of an electron diffusion length. The recombination luminescence so produced may be of sufficiently high intensity. Light is emitted at small direct voltages of about several volts (slightly more than the potential difference conformable to the band gap width of the material in which the *p-n* junction is formed).

The wavelength of recombination radiation exceeds somewhat the wavelength corresponding to the band gap width of a semiconductor material. As a rule, the emitted light exhibits good monochromaticity—the peak width of the spectral distribution of recombination radiation lies within 200 to 600 Å. The emission spectra obtained at present (depending on the starting semiconductor material used) extend from a far infrared (for InSb) to a green and dark blue light region (for SiC). The emission of light is easily modulated by the current of the light-emitting diode and has a high speed of response from 10^{-6} to 10^{-9} s. The service life of these diodes may be brought up to 10 thousand hours. Best results in the emission intensity are attained at present with the use of fight-emitting diodes made

from the compositions of the type $A^{III}B^{V}$ (GaAs, GaP, InSb) and their solid solutions ($GaAs_xP_{1-x}$, $GaAs_xAl_{1-x}$) and also from SiC.

The emission intensity of light-emitting diodes is increased by their clarification (covering the diode surface with a transparent dielectric film of a thickness in the order of one fourth of the emitted light wavelength) and also by making the diodes hemispheric in shape. In the diodes of such a shape the emitted radiation is incident on the interface at right angles to almost all its points and is subject to least scattering.

Despite its good monochromaticity, the recombination radiation in conventional injection-type light-emitting diodes is incoherent, its intensity is not particularly high and the efficiency comes to merely several per cent. The incoherency of radiation means that every act of radiative recombination is independent of others and light wave phases are different.

Light emission may be significantly intensified through the generation of coherent radiation.

If a system consisting of excited and nonexcited levels is in the state of thermal equilibrium, the number of electrons at lower levels is always greater than that at higher levels (the law of distribution of carriers over energy levels is defined by the Fermi function). In this case a continual exchange of electrons takes place between the higher and lower levels in the atom (or between the conduction and valence bands of the semiconductor), i.e., the electrons are excited from the lower levels to the higher levels and vice versa. The first process involves the absorption of light and the second, emission of light.

The probability of this or other process depends on the number of electrons distributed over the energy levels. This is why in a solid body in the state of thermal equilibrium the probability of light absorption which is in proportion to the number of electrons occupying lower levels is higher than the probability of light radiation.

Assume now that the number of electrons on the higher level is larger than the number of electrons on the lower level. This state, termed inverse population, can be created in a solid body (and also in a molecule of gas or liquid) through intensive excitation, also called pumping, of electrons from the lower levels to the levels of higher energy.

The pumping of electrons in a semiconductor body can be carried out by optical means, by collision ionization in a high electric field or by injection of electrons and holes through a *p-n* junction.

The inverse population of energy levels is usually created just near the bottom of the conduction band and near the top of the valence band (electrons move from the higher levels of the valence band to the lower levels of the conduction band).

If a solid body with the inverted population of energy levels is irradiated by the light with a quantum energy corresponding to the distance between the bands, the light will not be absorbed.

On the contrary, the light falling on the body causes radiation rather than absorption because the radiation probability being proportional to the number of electrons in the conduction band will be higher than the probability of absorption. The unitary process of interaction of an incident photon and an excited electron is such that the photons ejected by the electrons returning to the lower levels become in phase with the photons that have caused this transition. In this way the incident radiation is intensified and becomes coherent.

Since the number of electrons on the energy levels of the conduction band and the valence band sharply diminishes with the distance from the band edges, the radiation here is monochromatic.

In order to utilize the coherent radiation so produced to the utmost and also to obtain an oriented beam of light, use is made of the operating principle of an optical resonator which provides positive feedback essential for the initiation of generation. Besides, the resonator draws quanta of light propagated in one direction.

To form a resonator, the lateral faces of a diode or crystal are disposed normal to the plane of a *p-n* junction. The faces are then polished to ensure good reflection.

In such a resonator the light repeatedly passes through the substance and gets intensified still further. The requisite condition for the intensification of the reflected wave, i.e., the coincidence (or the difference by an integral number of periods) of the direct wave phase with the reflected wave phase, is achieved by selecting a required distance between the reflecting faces, in other words, by choosing a proper length of the optical resonator.

In semiconductor optical quantum oscillators (lasers) the inverted population is created by irradiation with an electron beam in vacuum or by the injection of electrons across a *p-n* junction. A high density of electrons on the inverted levels (in the order of 10^{18}-10^{19} cm^{-3}) is obtained by passing a heavy direct current through a *p-n* junction in which both layers are degenerate, i.e., the Fermi level lies in the conduction band and in the valence band (as in the tunnel diodes).

Such a *p-n* junction may be used as an amplifier though it usually serves as a light generator.

When a sufficiently high current in excess of a certain threshold value is passed through such a diode, a progressively intensified process of light generation and amplification arises in the diode, this process being balanced out by the losses due to the radiation of light into the air medium.

Significant achievements have been made in the construction of semiconductor lasers on this principle. They operate in the con-

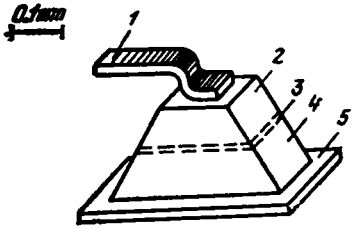


Fig. 10.15. Semiconductor laser structure
1 and *5*—electrodes; *2*—*n*-type GaAs; *3*—*p-n* junction region; *4*—*p*-type GaAs

ditions of continuous generation (continuous wave lasers) at low temperatures and in the conditions of pulsed generation (pulsed lasers) at room temperature.

Given below are the wavelengths for semiconductor injection lasers classified by the basic materials they are made from:

gallium arsenide	0.85 μm
indium phosphide	0.9 μm
lead sulphide	4.3 μm
lead telluride	6.5 μm
gallium arsenide-phosphide	0.65-0.9 μm
indium arsenide-phosphide	0.9-3.2 μm

Lasers based on arsenide and gallium arsenide-phosphide have found most extensive use. Thus, gallium arsenide pulsed lasers operating at room temperature have an emitting power of 10 to 12 W at a pumping current of 90 A and an efficiency of 15 per cent, the duration and repetition frequency of pulses being 100 ms and 10 kHz.

The power and efficiency of a laser may be increased by lowering its temperature. For example, at 77°K (the temperature of liquid nitrogen) a 40 per cent efficiency can be reached at a pumping current of merely 100 mA. The arrangement of a gallium arsenide laser is illustrated in Fig. 10.15.

Review Questions

1. What causes maxima in the spectral sensitivity and intrinsic photoconductivity?
2. How is the type of recombination related to the illumination intensity?
3. What type of recombination prevails in a heavily doped semiconductor?
4. How can one determine the carrier lifetimes by the shape of the photocurrent build-up curve?
5. In what way is it possible to expand the semiconductor photosensitivity spectrum?
6. What makes the photoresistor differ from the photodiode?
7. What is the difference between photocells and photodiodes?
8. Why does the phototransistor sensitivity exceed that of a photodiode?
9. Why should the *p-n* junction of a photodevice be disposed near the crystal surface?
10. Why are the emission spectra of light-emitting diodes generally narrower than the electroluminescence spectra?
11. What uses of lasers do you know?

Chapter Eleven

INTEGRATED CIRCUITS

11.1. RELIABILITY OF SEMICONDUCTOR DEVICES

By the reliability of any device is meant the probability that the device will perform its function for a specified interval under stated conditions. The reliability of a semiconductor device is characterized by the probability of no-failure operation within a previously specified period of time. The failure is a change in the device parameters, which leads to a disruption of normal operation of a circuit or a shift of these parameters beyond specified limits.

Failures are subdivided into sudden (catastrophic) and gradual (incipient) failures. The occurrence of sudden failures results in complete inability of a device to perform its function. As a rule, these failures are attributable to shorts or breaks of internal or external leads in the device or a breakdown of the *p-n* junction. These failures are also called unconditional as they render unserviceable any circuit that contains a disabled component.

Incipient failures lead to gradual changes in the parameters of a device. The magnitude and rate of these changes can be different, and when the parameters change so fast that the device itself fails to perform its function, the incipient failures tend to turn into catastrophic ones by their manifestation. Yet, most often these changes come about progressively in such a manner that the operational ability of the device fades out long before the moment the device is made fully incapacitated. The influence of changes in the parameters of a device on its operational ability is naturally accounted for by the type of the circuit and for this reason the failures of this kind are called conditional.

To control the quality and check the reliability of products turned out by the manufacturer, they are put to tests of various types.

It should be remarked in passing that it is most difficult to define a universal criterion for a conditional failure since any testing is performed in accordance with a certain scheme and does not make it possible to apply the test results obtained by this scheme in any other test programmes.

It is essential that all the tests should be run for short intervals because prolonged tests do not allow timely corrections to be introduced into the production process.

The routine testing usually includes mechanical endurance, climatic and life tests. Mechanical endurance tests help determine the resistance of a device and its terminals to the stresses encountered in operation: shocks, vibration, etc. Climatic tests define the in-

fluence of temperature, pressure and humidity variations on device parameters. On trials the devices are generally subjected to sharp temperature differences (thermal cycling). Life tests are conducted on large batches of devices over 100, 500, 1000 hours or even longer under specified conditions.

By the number of the devices failed during the tests one can determine such an important reliability parameter for semiconductor devices as failure density (rate) λ.

The parameter λ is determined experimentally from the formula

$$\lambda = \frac{C}{nt} \tag{11.1}$$

where C is the number of semiconductor devices that have become inoperative for time interval t, and n is the number of the samples which have withstood the tests during test time t in hours.

Another reliability parameter is the probability of no-failure operation of a device for a period of time t. This parameter is related to the failure density parameter by

$$P(t) = 1 - \lambda t \tag{11.2}$$

In semiconductor devices unconditional failures account for 60 per cent of all failures.

Conditional or gradual failures most often ensue from changes in the condition of a semiconductor surface. Indeed, owing to sorption of oxygen and moisture by the surface of a semiconductor, surface charges, and also a surface potential and conductivity vary to an appreciably large degree.

The presence of charges on the surface causes the appearance of conducting channels between p- and n-regions, sharp variations in the breakdown voltage (see Sec. 2.6) and other phenomena worsening the characteristics of p-n junctions.

The process of changes in the surface charge may continue for a very long time, tens and hundreds of hours, and during this period the characteristics of devices and circuits become gradually altered. The planar technique ensures a more stable state of the surface, for in this case the surface is protected by an oxide film. But the presence of ions of alkali metals which vigorously diffuse across the oxide film gives rise to time-variable leakage currents in the p-n junction and other undesirable effects. Thus, the diffusion of positive ions of alkali metals through the oxide film results in a built-in channel in a silicon induced n-type channel MIS transistor. This happens because at the interface between the insulator and the silicon layer an induced charge of electrons appears in the silicon, which cancels out the positive charge of ions in the insulator. This charge of mobile carriers forms a conducting channel between the source and the drain.

When semiconductor devices make up a circuit containing N elements of k types, the failure density of the circuit is calculated by the formula

$$\lambda = \sum_{i=1}^{k} N_i \lambda_i \tag{11.3}$$

It follows from the formula that with an increase in the number of components of the circuit, the reliability of the latter decreases.

Research data show that over 50 per cent of failures of radioelectronic apparatus are associated with the failures of contacts and therefore the number of failed electronic circuits grows with an increasing number of component connections.

The density of circuit failures associated with the number of failed components and contacts and the number of components and contacts in the circuit is expressed in the form

$$\lambda_p = \sum_{i=1}^{l} \lambda_{ci} N_{ci} + \sum_{i=1}^{k} \lambda_{Ci} N_{Ci} \tag{11.4}$$

where λ_{ci} and N_{ci} are the density of failures of contacts and the number of contacts, respectively; and λ_{Ci} and N_{Ci} are the density of failures of components and the number of components in the circuit, respectively.

The system assembled of a large number of separate components has a far lower reliability compared to the reliability of an individual component because of a large number of connections between the components.

With an increase in the number of components in circuits, methods that help reduce the size of radio apparatus containing tens and hundreds of thousands of elements acquire great importance.

It can be said with assurance that technical advances made in aeronautics, astronautics and other branches of science and engineering would be impossible without small-size electronic computers composed of semiconductor devices.

An important characteristic of an electronic apparatus is the packing density or the density of components per cm^3. This value for a vacuum-tube receiver is less than $10^{-2}/cm^3$ (10^{-2} components per cm^3).

Small dimensions of electronic devices can be attained by reducing the size of transistors, resistors, capacitors, diodes and also by elaborating printed-circuit assemblies and micromodule structures. Owing to this approach the circuit density has been brought up to $10^2/cm^3$.

This value is likely to be the upper limit in the so-called discrete miniaturization of electronic apparatus. Under this term is understood the reduction of dimensions of individual components of a

circuit. This limit is dictated by the component size which is so small that the assembly operations become very complicated and with the circuit density in excess of $10^2/cm^3$ they prove purely impracticable.

It is microelectronics that has made it possible to sharply raise the packing density, to lower the cost of electronic devices and increase their reliability.

11.2. MAJOR OBJECTIVES OF MICROELECTRONICS

The principal task of microelectronics is to increase the reliability and circuit density of radioelectronic apparatus.

One of the ways of increasing the reliability is a reduction in the number of connections between the components of the circuit. For this purpose such technological processes are used in which the production stages for a component are combined with those for a circuit as a whole. For example, the planar technique makes possible a *simultaneous* preparation on a single crystal of a transistor, diodes, a resistor and connections there between all of them making up the circuit shown in Fig. 11.1. In this circuit the diffused layer of a *p*-type semiconductor serves as a resistor, while contact area *1* connects the resistor with two diodes and contact area *3* connects one of the diodes with the base of the planar transistor. As is shown in the drawing, the active components are made by the same methods as are used for the preparation of discrete transistors. The only diffe-

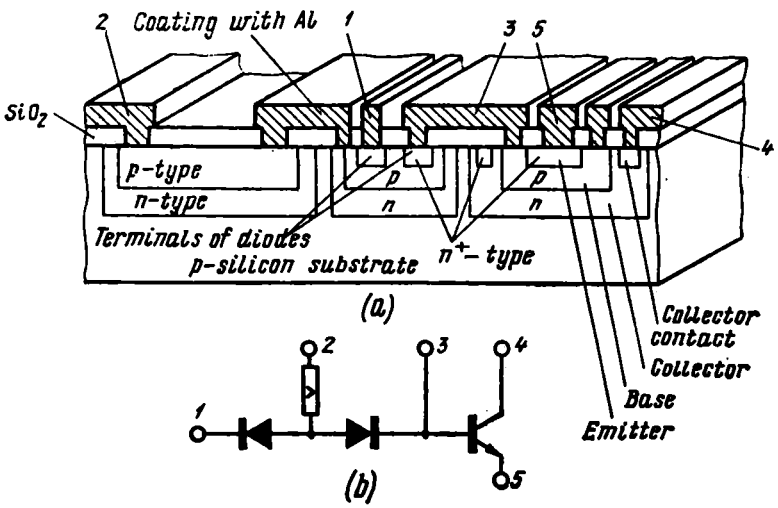


Fig. 11.1. Semiconductor IC
(a) sectional view of the structure; (b) electric circuit

rence is that by the planar technology intercomponent connections are made concurrently with the preparation of the active elements of the circuit in one and the same crystal.

As circuit components are isolated from each other mainly by a reverse-biased *p-n* junction, with a great packing density spurious couplings may arise between the adjacent elements, which limit further growth in the density of components within a single semiconductor crystal. The technology that unites the stages of manufacture of a component and the whole circuit is called integrated-circuit technology and the circuits made by this technology are called integrated circuits (IC).

The integrated-circuit processing technology has made it possible to cut down the number of contacts in the circuits of the same type by 30 per cent (by 4 contacts out of 9 in the circuit shown in Fig. 11.1*b*); furthermore, the reliability of such integrated circuits increases almost tenfold as against the circuits made up of discrete components. Both bipolar and MIS transistors are used as active components of integrated circuits which are respectively identified as bipolar ICs and MIS-ICs.

Owing to the achievements in the integrated-circuit technology, the reliability of semiconductor integrated circuits containing tens of active components is at present equal to, or in some cases even higher than, the reliability of individual discrete transistors.

Since the IC production process dispenses with soldering and welding, used to provide connections between components, it has become possible to markedly reduce the size of components. Thus, MIS transistors incorporated in ICs turned out by the semiconductor industry measure each about 20 by 20 μm. One semiconductor wafer measuring $4 \times 4 \times 0.2$ mm^3 can accommodate one thousand circuit components. The packing density or the number of circuit components per unit volume of a semiconductor amounts to more than 10^7/cm^3. The trend today in the development of microelectronics is toward still further increase in the packing density. The man's brain can probably serve as an example of the most dense packing of components. Taking up the space of 1.5×10^3 cm^3, the brain contains 10^{13} components (neurons), which corresponds to the packing density of 10^{10} components per cm^3. Every neuron dissipates a negligibly small power of $(2\text{-}8) \times 10^{-9}$ W/cm^3 or 10^{-10} W per component.

The packing density of the available integrated circuits is by far lower and the dissipated power higher as compared to these values for the man's brain, but with each passing year the circuit density grows and, as has been already stated, has reached 10^6-10^8/cm^3.

Much consideration has to be given to the problem of reducing the power dissipated by components. An increasing packing density and thus better compactness of a device result in a growing rate of

response of the circuit due to a reduction in the propagation time required for a signal to cross a path from one part of a circuit to another (the delay of a signal covering a path 33 cm long is 1 ns). Of particular importance is the reduction of dimensions of the circuits incorporated in microwave devices.

In existing integrated circuits, the component density per crystal has been brought up to $10^6/cm^3$. This quantity in packaging integrated circuits is about a hundred times as low.

The packing density in complex units composed of several integrated circuits becomes still lower on account of the space taken up by commutating elements. This example is illustrative proof of the need of converting to large-scale integrated circuits (LSIC) which contain one thousand components and even more in one crystal. A quantitative assessment of the packing density of IC is given by the level of integration g, which is the logarithm of the number of components entering into IC

$$g = \log N$$

Thus the IC containing one thousand components is identified as the IC of the third degree of integration.

Limitations on the packing density of components within one crystal may be subdivided into the following groups.

1. Limitations dictated by technological possibilities.

Only a few years ago it was believed that the elements of less than 3 μm in size were impossible to obtain, for at smaller sizes inaccuracies in the production of photomasks caused appreciable spread of geometrical dimensions and parameters of circuit elements. Now, owing to a high resolution of optical devices and photoresists it has become possible to prepare MIS structures with a channel 0.8 μm long, though, the channel width comes to tens of microns. Since a rise in the level of integration is reached both by increasing the area of a crystal and by diminishing the size of an element, severe limitations are imposed by a restricted resolution of the optical apparatus used in photolithography. Therefore, it is intended in the future to process the photoresist with an electron beam which has a lesser diffraction than the light and permits achieving higher accuracy over a larger area.

2. A nonuniform distribution of impurities in a crystal or semiconductor film has to be reckoned with at small sizes of a component.

3. A high density of components necessitates accounting for the power dissipated in IC. Indeed, if in a circuit comprising 30 transistors and occupying a space of 10^{-3} cm^3 each transistor dissipates 1 mW, its specific dissipated power totals 30 W/cm^3. The quest for a decrease in the dissipated power per component of a circuit becomes thus quite understandable.

At present, MIS structures and particularly circuits composed of transistors with complementing types of conduction allow for decreasing this power to tens of microwatts. The next level to be won over by microelectronics is a nanowatt range of power. The main way of reaching this level lies in reducing the power consumed and dissipated by the transistor in a steady condition. Let us note that in bipolar-transistor integrated circuits the dissipated power (5-50 mW) is many times higher than that in MIS-ICs (10^{-6}-5 mW).

4. Limitations on the density of components within a crystal and a substrate are also due to spurious couplings. In semiconductor integrated circuits (an example of such a circuit is given in Fig. 11.1), these couplings are produced by currents and capacitances of reverse-biased *p-n* junctions which isolate one component from another. A large density of components disposed between the *n*-regions separated by the *p*-areas (or vice versa) may give rise to a transistor effect that causes unwanted couplings between the components. Besides, the capacitances of decoupling *p-n* junctions operating at high frequencies can also produce spurious couplings.

Spurious couplings are less pronounced in the so-called film circuits. Here they arise mainly from the capacitances of supply busbars and contact areas.

The comparison of a conventional circuit with a microelectronic circuit serving the same purpose shows that though the size, weight and cost of the microcircuit are considerably smaller and the reliability is higher than the same parameters of the conventional circuit, the number of elements remains the same. Since an ever growing sophistication in the design of the devices is one of the main tendencies in the development of microelectronics, the problem of arrangement of an increasing number of elements and their interconnection becomes more and more acute.

It is not difficult to foresee that under these conditions a new quantitative problem will emerge in the future, the solution of which with the aid of the microelectronics of "discrete" elements is impossible. Really, if the rate of failure of a component is as low as 10^{-7} 1/h, the life of a circuit (without redundancy) made up of over 10^7 components, which will be undoubtedly achieved soon in practice, can be only one hour.

In connection with this, of much interest is a search for ways of evolving the devices that will be able to fulfil a certain function in a circuit, as of a generator, amplifier, converter, etc. The salient feature of these circuits is the fact that it is impossible to single out here particular components serving the role of a transistor, resistor, capacitor, etc.

An example of such a functional device is a piezoelectric delay line which comprises a quartz crystal with two piezoelectric transducers (Fig. 11.2*a*). The electric signal passing through the delay

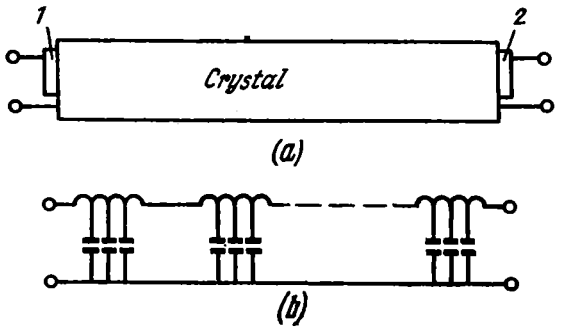


Fig. 11.2. Piezoelectric delay line
(*a*) arrangement; (*b*) equivalent circuit; *1*—input piezoelectric transducer; *2*—output piezoelectric transducer

line is held over for time t dependent on the size of the quartz crystal.

The delay line depends for its operation on the difference between the speed of propagation of a signal (velocity of light) and sound velocity in dielectric. An input piezoelectric transducer converts electric oscillations to sound waves which travel through the quartz crystal with the speed of sound, reach an output transducer and are transformed there into an electric signal.

An analogous conventional circuit includes several tens of series-connected LC paths (Fig. 11.2*b*).

Functional devices can probably become the basis of most of complex radioelectronic units.

11.3. MAIN PRINCIPLES OF PREPARATION OF INTEGRATED CIRCUITS

We consider now the main methods of manufacture of ICs.

Semiconductor integrated circuits are made by the methods of selective single or multiple diffusion of impurities into a semiconductor monocrystal, forming thereby in its certain portions diodes, triodes and passive components. Their interconnection is done either directly within the bulk of the semiconductor or by means of special selective coating of the crystal surface with metal. Semiconductor integrated circuits made in a single semiconductor crystal are called *monolithic*.

An example of such a monolithic integrated circuit is illustrated in Fig. 11.1.

Film integrated circuits are prepared by the method of sequential deposition on the insulating substrate of thin metal, dielectric and semiconductor films of certain shape, which thus form all the circuit elements.

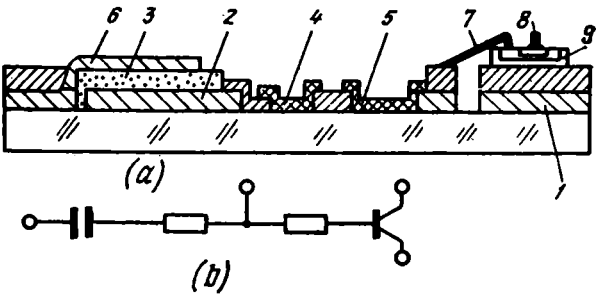


Fig. 11.3. Hybrid IC

(*a*) disposition of film module components; (*b*) electric circuit of the film module; *1*, *2*, *6*—aluminium films; *4*, *5*—high-resistivity film forming resistors; *3*—capacitor dielectric; *7*—base; *8*—emitter; *9*—collector

A serious restriction on a wider use of film ICs is the absence of a film transistor and therefore only passive components (resistors and capacitors) are made by the deposition of films on the substrate. Caseless miniature transistors are then attached by various methods to specially prepared contact areas. These circuits are termed *hybrid.* Their component density is comparatively low—only several hundreds per cubic centimetre, but the speed of response of these circuits can be rather high due to shortening of connections. A reduction in cost and size of passive components can ensure the market for these circuits.

Figure 11.3 shows a hybrid circuit built up of a transistor, two resistors and a capacitor, and also an arrangement of circuit components on the substrate.

On the insulating substrate made, for example, of glass or pyroceram are deposited aluminium film *1* forming contact areas and lower plate *2* of the capacitor. The next step includes depositing dielectric layer *3* for the capacitor and a film of a metal or an alloy of high resistivity to form resistors (*4* and *5*). Further steps comprise the preparation of upper plate *6* of the capacitor and gold contact areas for a caseless transistor to be attached thereto by soldering.

The description of the production process is certainly made in a very simplified manner. In actual fact, chrome or tantalum films are deposited underneath the gold or aluminium contact areas to enhance the cohesion of aluminium or gold with the substrate. The dielectric layer can also be formed by oxidizing the lower capacitor plate (aluminium or tantalum).

All integrated circuits are made flat and for this reason their packing density is determined by the area a component occupies on a crystal or substrate surface. The notion of surface resistivity or resistance expressed in ohms per square is introduced to describe resis-

tors and conductors. Indeed, if the length and width of a conducting element are equal, then

$$R = \rho \frac{l}{lh} = \rho \frac{1}{h}$$

where h is the thickness of the film.

The resistance of a square area of any size depends thus only on the resistivity of the bulk of the material and its thickness. The concept of capacitance per unit area of the surface, measured in pF/cm^2, is likewise used to characterize capacitors.

It is obvious that the higher the nominal values of resistance and of capacitance, the larger the area a component must occupy on the surface of IC. In integrated circuits prepared from semiconductor with a relatively low resistivity, resistors take up a much larger area than transistors. The same can also be said about capacitors. Since both resistors and capacitors are formed in a semiconductor crystal by the same methods as are transistors and diodes, the comparative cost and density of components in the circuit are determined by their surface area. So the fabrication of semiconductor ICs with passive components is less profitable than the fabrication of the circuits that include only active components, because they occupy by far greater an area than the transistors and their comparative cost is much higher than the cost of the transistors.

The trend today is to develop circuits composed completely of active components. When it is necessary to preserve passive elements, hybrid circuits are made by the method of combined technology.

These circuits are made on the same principle as are the hybrid circuits with discrete microtransistors but the transistors that are attached to contact areas of passive portions of the circuits are replaced here by semiconductor circuits. In this way the film and the integrated-circuit technology are combined to produce these circuits.

Thus, hybridization in the production of large-scale integrated circuits manifests itself in that the connection between semiconductor integrated circuits is by means of layers.

Another variant of the combined technique is the deposition of passive components on the oxide layer that protects a semiconductor integrated circuit.

11.4. SEMICONDUCTOR INTEGRATED CIRCUITS

The semiconductor integrated circuit is a semiconductor monocrystal, individual portions of which are used to perform certain functions. Circuit components, such as diodes, transistors, resistors and capacitors are prepared by the planar or planar-epitaxial growth technique.

In these circuits the barrier capacitance of a *p-n* junction is used to form capacitors, and for this reason methods of production of dio-

des and capacitors are identical. The MIS structure is sometimes used as a capacitor. In both cases the required value of capacitance is achieved by selecting a proper level of doping of the semiconductor and a required area for the diode.

Electrical separation between two components is most often provided by a reverse-biased *p-n* junction. Fig. 11.1 shows how resistors, diodes and transistors are isolated from one another by *p-n* junctions. To improve isolation of components, a blocking voltage is sometimes impressed across the substrate. Since the resistance of the back-biased *p-n* junction in silicon is as high as hundreds of megohms, this type of insulation proves quite reliable for many circuits. Insulation of components is occasionally effected with the aid of the so-called "pocket" technique by which components are isolated from one another with specially grown dielectric regions.

Resistor width d in semiconductor ICs cannot be smaller than the depth of a diffused layer and therefore its size ranges from 5 to 10 μm. But resistor length l must be appreciably great and thus a resistor with a resistance of several kilohms occupies a large area on the crystal surface, which considerably exceeds the area of a transistor (20×20 μm^2). Integrated circuits containing passive components usually include bipolar transistors serving as active components. Such circuits can exhibit a sufficiently high speed of response. The switching time of these circuits lies in the range of 5 to 20 ns and the dissipated power, from tens of milliwatts to several microwatts.

But the level of integration in these circuits is not high because large dimensions of transistors and passive components do not as yet allow for the allocation of more than 50-100 components in one crystal.

A far greater level of integration (higher than the fourth degree at the present time) is achieved in MIS circuits because MIS transistors take up 5 to 10 times as small an area on the crystal surface as bipolar transistors and, besides, they have a higher noise immunity (due to a clear-cut trigger level). Unlike bipolar ICs, MIS-ICs have the drain and the gate supply voltage of the same sign. This obviates the use of matching components which are indispensable in bipolar transistor circuits.

MIS circuits dispense with passive components for a MIS transistor is used as a load resistor, while the load resistance is governed by the voltage across the gate and the drain. Transistors in MIS-ICs are isolated from each other by back-biased *p-n* junctions. The percentage yield of these circuits is high. A relatively high scrap rate in the process of production of bipolar ICs is due to the fact that a high-frequency transistor requires a very thin base of about one micron.

The base width depends on the temperature and the time of diffusion and hence a thorough control over the production process is

imperative. At the modern level of semiconductor technology the scrap rate attributed to inaccuracies in the production of very thin bases is impermissibly high. The diffusion depth in MIS transistors is not so critical and the percentage yield of serviceable circuits is therefore higher.

MIS circuits have found especially wide use in computing machinery. The level of integration of MIS-ICs grows unabatedly. The size of components will probably reach soon the limit conditioned by the resolution of the photomasking process (wavelength of light, diffraction) and the accuracy of optical devices. In this connection, growing complexity of circuits leads to an increase in the crystal area occupied by IC. This in turn causes an increase in the percentage waste due to the defects whose probability of occurrence on the working region of IC grows with a rise in its area. These defects result in puncture of oxide films and breakdown of *p-n* junctions.

A decrease in the size of MIS transistors and increase of their response are realized through the methods of self-coincidence (silicon gate, molybdenum gate, ion implantation) described in Chapter Eight. Designing of integrated circuits with a high level of integration involves elaborate calculations, for which purpose electronic computers are used.

Computer-aided design of integrated circuits helps calculate and draw for a short time photomasks of complex ICs containing thousands and tens of thousands of components.

Computers are extensively used in testing ICs according to prescribed programmes, enabling thereby a manifold decrease in the time and labour required to perform control operations which account for an appreciable part of expenses in the production of ICs.

11.5. FILM INTEGRATED CIRCUITS

Let us examine the methods for preparation of thin films and film elements.

The surface of a solid body is always subjected to some pressure of saturated vapours. At room temperature the pressure of saturated vapours of the materials used in microelectronics reaches 10^{-8} to 10^{-20} mm Hg and therefore the evaporation of the materials in these conditions may be disregarded. If the ambient pressure is reduced below the saturated vapour pressure, a substance starts evaporating.

Evaporation of a substance is commonly conducted by heating it in a vacuum at a pressure of 10^{-5} to 10^{-9} mm Hg. The temperature of the substance being evaporated is brought up to such a value that the vapour pressure greatly exceeds the pressure of the surrounding medium. Evaporation is believed to proceed most intensely

at a temperature at which the vapour pressure usually makes up 10^{-2} mm Hg.

The temperature dependence of pressure may be described by the empirical relation:

$$\log P = -\frac{A}{T} + C \tag{11.5}$$

where A and C are the constants of a substance.

The values of constants A and C for some metals are listed in Table 11.1

Table 11.1

Constants A and C for Metals

Metal	Constants		Metal	Constants	
	A	C		A	C
Cu	12,300	7.445	Cd	6073	11.720
Ag	12,270	8.39	Al	17,250	9.705
Au	12,450	8.028	Si	20,615	10.730
Mg	7167	8.088	Pb	9870	7.85
Zn	6750	8.916			

The evaporation temperatures are rather different for various substances (with $P = 10^{-2}$ mm Hg); thus, selenium evaporates at 217°C and tungsten at 3382°C.

Such a great difference in the evaporation temperatures permits using tungsten, tantalum ($T_{ev} = 2996$°C), carbon (3700°C) and some other high-melting metals for constructing heaters — evaporators intended for vapourization of substances. There is a great variety of evaporator designs.

Thus, a resistance evaporator is heated when the electric current is passed through it. In electron bombardment evaporators the heating of the evaporating substance is effected by an electron beam emitted from a tungsten cathode disposed around a bar of the evaporating substance which is connected to a positive terminal of the high-tension source. The voltage of 3 to 5 kV impressed across the cathode and anode turns out to be sufficient to evaporate even such refractory metals as tantalum and tungsten. The electron bombardment process excludes a contact between the evaporating substance and other materials in the evaporation zone and thereby this method allows for obtaining most clean films devoid of the metals of the evaporator.

In plasma evaporators the evaporating substance is heated by ions of an inert gas fed into the setup. An example of plasma evaporation may be cathode sputtering. The substance to be evaporated serves as a cathode whose surface is bombarded by the ions produced

in the gas discharge between the anode and the cathode at a voltage of 1.5 to 3 kV and a gas pressure from 10^{-1} to 10^{-3} mm Hg. As the evaporation rate is proportional to the energy of the ions bombarding the surface, the cathode evaporation is carried out in the atmosphere of heavy inert gases, for instance, argon. The inert atmosphere is needed to prevent oxidation of the substance by the oxygen ions occurring in the residual gases.

While evaporating, the substance vapours spread in vacuum uniformly in all directions. The rectilinear propagation of molecules becomes possible if the vapour pressure does not exceed 10^{-2} mm Hg. Under these conditions the particles are regarded to travel without collisions.

If in the way of the molecular flow a substrate is placed, the surface temperature of which is below the evaporation temperature, the substance will begin to condense and deposit on it. Condensation rate ν varies with the temperature of the condensing surface by the law

$$\nu = e^{Q_{ad}/kT}$$

where Q_{ad} is the adsorption energy defining the cohesive force between the atoms of the evaporating substance and those of the condensing surface, and T is the temperature of the condensing surface.

As a rule, the higher the melting point of a substance, the larger the value of Q_{ad}, and the stronger the atoms of the evaporating substance are bound with the atoms of the condensing surface.

The film circuit substrate must comply with a variety of requirements, the main ones being strength, good insulating properties, optical surface, chemical inertness and resistance to high temperatures.

Extensive use has been made of such materials for substrates as glass, pyroceram, quartz and mica.

The requisite shape of film elements is achieved by spraying film substances through special masks tightly pressed to the substrate surface.

The size of film elements is generally very small and therefore the holes in the mask must measure microns. These holes in metallic masks are obtained through photomasking or chemical etching. The preparation of masks measuring microns in size is a fairly complicated task, the more so when the deposition of films is performed at elevated temperatures causing thermal deformation of substrates and masks. Under development now are the methods for the fabrication of films of certain shape by photolithography. The resolution attained in this process lies in the range of several microns, but the photoprocessing technique used for the preparation of patterns in multilayer circuits requires further improvement.

At least three processes of spraying are required in manufacturing a passive or an active film circuit: deposition of low-resistivity metals to provide intercomponent connections and contact areas, of high-resistivity metals or alloys to form resistors, and of dielectrics for capacitors.

Should the circuit contain active components or there is need for a multilayer arrangement of circuit components, the number of operations is increased correspondingly.

Film circuits are prepared in multistage vacuum setups which allow performing 6 to 8 operations of spray-coating of several substrates with different substances during one cycle of vacuum deposition. The setup is usually provided with two revolving tables one of which carries evaporators and the other, masks and substrates. Performing several operations of spraying for one cycle of gas evacuation helps achieve a higher productivity and a lower cost of film circuits.

The nominal resistance values of *film resistors* are determined by their shape, resistivity of a metal or an alloy they are made from and a film width. Resistors with a low resistance from 200 to 500 ohms are made from films of metals and alloys, most commonly chrome and nichrome (80% Ni, 20% Cr). It would seem possible to increase the resistance of a resistor by reducing its thickness h, but already at $h = 100$ Å the film ceases to be a continuous layer and consists instead of separate poorly contacting small islands of a metal or alloy on the substrate. As the resistance of such films is not stable, the thickness of film resistors is made to range from 200 to 700 Å.

To obtain resistors with a resistance of 1000 ohms and even more per square of surface, use is made of metal-ceramic films which are prepared by depositing a metal-dielectric mixture on the substrate.

Thus, the Cr-SiO composition, called cermet, is used to prepare resistors with 250 to 3000 ohms per square depending on the chrome content in the composition.

The operational stability of resistors is improved by subjecting the deposited films to annealing. This process causes recrystallization and oxidation of the films, owing to which their parameters become more stable.

Resistive films are commonly deposited at the temperature of the substrate reaching several hundred degrees C and therefore resistors are often arranged in the first layer of the film circuit.

The temperature coefficient of resistance (TCR) of a film resistor can vary substantially depending on the composition of the resistive film. Thus, for the cermet composition including 70% Cr and 30 % SiO, the TCR reaches 5×10^{-6} 1/°C.

Film capacitors are made from the films of various dielectrics: SiO, SiO_2, Al_2O_3, Ta_2O_5, etc. Some films, for example, Ta_2O_5 and Al_2O_3, are prepared by oxidizing appropriate metal films, the others, as silicon monoxide SiO, oxides of rare-earth metals (dysprosium oxide Dy_2O_3, for example) and borosilicate glasses are obtained through evaporation in vacuum.

Film capacitors manufactured at present with a dielectric thickness of 2000-20,000 Å have a capacitance in the range of 0.001-0.01 $\mu F/cm^2$ and a breakdown voltage of up to 100 V.

The capacitance, dielectric loss tangent and stability of film capacitors heavily depend on the conditions of evaporation. An increase in the evaporation rate leads to a partial decomposition of oxide attended with liberation of a pure metal. This results in an increase of the dielectric conduction and of the loss tangent of the capacitor.

The conditions of spray-coating (evaporation and condensation temperatures) can also alter the distribution of mechanical stresses in the film, which appreciably affect the capacitor stability. The production process is therefore conducted under the conditions which cause minimal tension in films.

Capacitors based on Ta_2O_5 and Al_2O_3 films are prepared by the methods of electrolytic or plasma anodization of metals; in the latter method the aluminium film is placed between the cathode and glow discharge anode in the oxygen atmosphere. The thickness of the oxide film is governed by the voltage applied to the electrolytic bath, and in the case of plasma anodization by the oxygen pressure and voltages across the anode and the sample being oxidized.

The method of anodization of tantalum films is convenient in that by varying the film thickness of an oxide and a metal one can obtain resistive, conducting and dielectric films from the same starting metal.

The film capacitor prepared from a Ta_2O_5 film 2400 Å in thickness has a capacitance of 0.1 $\mu F/cm^2$, working voltage of up to 50 V and $\tan \delta = 0.01$.

One more technique of fabricating dielectric films consists in reactive deposition which represents the cathode sputtering of a metal in the oxygen atmosphere. By regulating the oxygen pressure and voltage it is possible to set up such operating conditions as to enable complete oxidation of metal molecules and condensation of an oxide layer on the substrate. This method offers good results in reactive sputtering of oxides of titanium, aluminium and silicon. The films obtained with the described methods have the thickness and the width of the lines ranging from 0.5 to 5 μm.

In a number of cases, when the high resolution is not needed and passive components account for a major part of the circuit cost, films more than 25 μm thick are used.

In the thick-film technique, the materials used for conductors, resistors and dielectrics are deposited on the substrate through masks prepared from special inks: a mixture of a glass and metal powder to obtain resistors and a dielectric powder with a binder for dielectrics. This technique is also called silk-screen printing because the pattern is "printed out" with ink on to the substrate through a silk cloth.

After drying, the substrates are placed in a furnace to burn out the binder and leave intact on the substrates sintered layers serving as resistors, conductors and insulators. The thickness of lines and their width reach approximately 25 μm and more. The cost of the circuits processed with the aid of the thick-film technique is nearly five times below the cost of thin-film circuits. But the lower resolution of thick films by comparison with thin films, and the difficulties encountered in obtaining homogeneous thick films make the thick film technique a poor substitute for the thin-film technology. The thick-film technique is of particular importance in the preparation of hybrid external commutation paths for semiconductor integrated circuits. The comparison between the parameters of passive components in thin-film circuits and thick-film circuits is given in Table 11.2.

Table 11.2

Comparison of Parameters of Thin-Film Components with Those of Thick-Film Components

Component type	Parameter	Circuit type	
		thin-film	thick-film
Conductors	Thickness of lines, μm Resistance, ohm/square	2-10 0.005-0.01	25 0.01-0.1
Resistors	TCR, % Resistance, ohm/square	±100 10-1000	100 1-200,000
Capacitors	Capacitance, pF/cm^2	23,000-5000	7500-15,000

Film portions of microelectronic circuits are usually composed of amorphous and polycrystalline films.

Film diodes can be prepared by forming either a blocking contact between a semiconductor and a metal (between gold and cadmium sulphide) or a *p-n* junction (between selenium and cadmium). The rectification ratios of such diodes are rather high (10^2-10^4), but the injection of minority carriers in polycrystalline-film *p-n* junctions

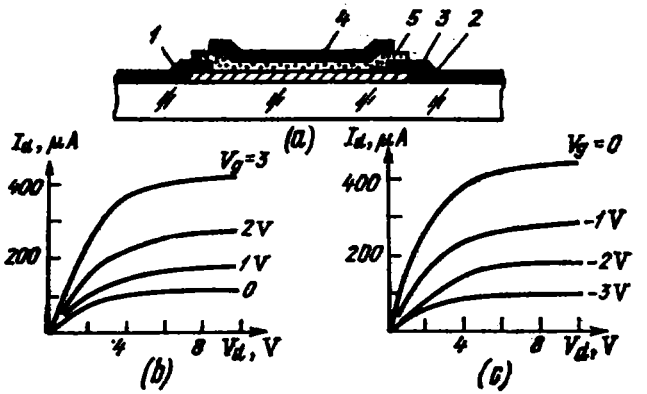


Fig. 11.4. FFET structure (*a*) and current-voltage curves of enhancement-mode type triodes (*b*) and depletion-mode type triodes (*c*)

1 and *2*—electrodes of the source and the drain; *3*—semiconductor film; *4*—gate electrode; *5*—insulator film

proves rather non-uniform over the surface area, which causes the generation of much noise. Besides, the low lifetime of carriers and the small diffusion length practically preclude the possibility of fabricating bipolar polycrystalline-film transistors.

The device whose properties are far less dependent on the crystalline structure than those of the bipolar transistor is a *film insulated-gate field-effect transistor*.

The operation principle of the film field-effect triode (FFET) resembles in almost every way that of the MIS transistor; the design difference lies in that electrodes *1* and *2* (Fig. 11.4) for the source and the drain respectively are ohmic contacts attached to uniform-conductivity semiconductor film *3*.

As a result, the current flowing through the channel of FFET in the absence of the voltage across the gate is not the current of the back-biased *p-n* junction but that which is determined by the conductivity of the semiconductor.

High-resistivity film field-effect triodes, such as cadmium sulphide film field-effect triodes, operate in the enrichment mode and low-resistivity tellurium-film triodes, in the depletion mode. A semiconductor film of medium conductivity, for example, a cadmium selenide (CdSe) film, is used for the fabrication of transistors that can be employed to function both in the enrichment and the depletion region.

The FFET, as any film element, is by far less sensitive to radiation than the MIS transistor. The transadmittance of the FFET is rather high and reaches 10 mA/V with a channel width of 1 mm. These

parameters of the triode permit its use in circuits of most general purpose. It should be noted, however, that to date all the advantages of the FFET still remain questionable owing to the insufficient stability of its parameters. Variations in the drain current arising from the changes in the state of the interface between the semiconductor and the dielectric under the influence of the voltage impressed on the gate account for 20 to 50%.

Changes in the characteristics with time (ageing) are typical of passive components too, though to a lesser extent. These gradual changes in the electric characteristics of resistive, dielectric and semiconductor films are in the main attributed to two factors.

First, changes in the resistance of films due to sorption of oxygen and water vapours. It is known that with the sorption of water vapours, the density of free electrons on the surface of an *n*-type semiconductor increases and with the sorption of oxygen, diminishes, in other words, water enriches and oxygen depletes the *n*-type semiconductor. Since sorption processes are appreciably inertial, the attendant changes in the parameters can last hundreds of hours.

Second, changes due to alteration in the crystal structure of films. Indeed, the films used for the fabrication of passive components are not monocrystalline but rather are polycrystalline and amorphous formations. These states of the structures prove thermodynamically unstable and in time, particularly at elevated temperatures, the amorphous films get crystallized and the polycrystalline films recrystallized. Changes in the size of crystallites in the films (recrystallization) bring about variations in the resistivity of the latter and alter the parameters of the devices and circuits incorporating these films.

The influence of gas sorption on film and semiconductor devices and circuits is usually lowered by sealing. Ageing caused by recrystallization may be slowed off by annealing the devices and circuits after their fabrication. Annealing speeds up recrystallization of materials and aids in stabilization of their characteristics.

11.8. EPITAXIAL-FILM INTEGRATED CIRCUITS

Intensive studies are under way now into devising integrated circuits that can combine the advantages of semiconductor integrated circuits and those of film circuits (insignificant spurious couplings). These circuits can be made up of monocrystalline semiconductor films (silicon, for example) deposited on insulating substrates.

Monocrystalline films permit the use of any method of planar technology for the fabrication of components. Besides, the photomasking technique can be employed to etch silicon in such a way that one component (a transistor or diode) can be completely isolated

from the other by the resistance of the substrate, while the interconnection of components can be provided by metal plating or selective etching as is done in semiconductor integrated circuits.

This technique rules out in principle the development of spurious couplings between components, which are inherent in semiconductor integrated circuits.

Monocrystalline films can be deposited on dielectric substrates by two methods: liquid-phase and vapour-phase epitaxy.

The liquid-phase epitaxial method consists in that prior to depositing a film the surface of an amorphous or polycrystalline substrate is coated with a low-melting glaze which liquifies at the deposition temperature and spreads into a thin layer over the substrate.

One of the compositions of the glaze used for the deposition of silicon contains

$$Na_2O - 1.8\%;\quad CaO - 7.1\%;\quad MgO - 2.5\%;\quad SiO_2 - 72.4\%$$

The substance particles condensed on the substrate form a monocrystalline film.

The monocrystalline layer with an ordered structure is formed because the condensate atoms acquire a sufficient mobility on the liquid substrate.

Vapour-phase epitaxy is carried out at high temperatures and therefore only those particles can hold on to the substrate which have penetrated into crystal lattice points. It is common practice to distinguish monoepitaxy when on the substrate is grown a monocrystalline layer of the same basic material as the original substrate, and heteroepitaxy when a monocrystalline film is grown on a single crystal of another material though identical in its structure with that of the film.

Monoepitaxy and, in particular, epitaxial growth of silicon on silicon, as has been mentioned earlier, is widely used in the fabrication of transistors and other semiconductor devices. Heteroepitaxy has found application in processing epitaxial-film integrated circuits.

To grow silicon single-crystal layers on the substrate of sapphire (α-modification of Al_2O_3) with the crystal lattice parameters close to those of silicon, use is made of the method of vacuum deposition in a deep vacuum (10^{-9} mm Hg) or the method of gas-transfer reactions. The epitaxial growth process is carried out in a deep vacuum (10^{-9} mm Hg) to prohibit oxidation of silicon during evaporation.

Silicon films with $n = 10^{16}$ cm^{-3} and $\mu_n = 300$ cm^2/V·s may be used to fabricate basically any semiconductor devices.

Since diffusion and particularly ion implantation permit obtaining *p-n* junctions at a small depth of penetration from the surface and thus forming thin base regions, it is possible to grow bipolar

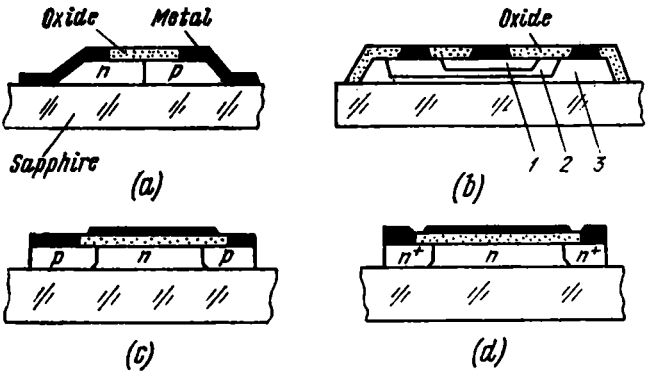


Fig. 11.5. Structures of components in epitaxial circuits (*a*) diode; (*b*) *n-p-n* transistor; (*c*) MIS transistor; (*d*) FFET; *1*—emitter; *2*—base; *3*—collector

high-frequency transistors in the film of only several microns in thickness.

Epitaxially grown films can be employed in the manufacture of other types of semiconductor devices too, namely diodes, MIS transistors and film field-effect triodes. The structure of these devices is illustrated in Fig. 11.5.

It is possible to produce transverse small-area *p-n* junctions in silicon epitaxial films. Indeed, with the width of the silicon strip of 20 μm and the film thickness of 1 μm, the area of the transverse *p-n* junction in a film diode measures 20 μm^2. The barrier capacitance C_{bar} of the junction is 0.005 pF.

With this method employed, it has become feasible to obtain on one sapphire crystal fast-response diode matrices (64 by 64 diodes) which find use in memory devices of electronic computers with a small time of sampling.

A severe disadvantage of epitaxial circuits is their high cost due to an extremely complicated process of growing and processing of the sapphire substrate.

Review Questions

1. Why is the level of integration of the circuits with active and passive components lower than that of the circuits made up of only active components?
2. Why does the reliability of integrated circuits exceed the reliability of discrete circuits of the analogous functional complexity?
3. What is the difference between the MIS transistor and FFET?
4. What physical factors limit the resolution of photolithography?
5. When and for what purpose is it necessary to decrease the power dissipated by circuit components?

Chapter Twelve

SEMICONDUCTOR TECHNOLOGY TODAY. TRENDS OF DEVELOPMENT

A large amount of research into the purification of semiconductor and other materials, creation of new techniques and perfection of previously known technological procedures, and elaboration of more sophisticated production and measuring equipment are responsible for an unprecedented advancement of the semiconductor industry.

The tendency for a rapid increase in the production of a growing variety of semiconductor devices due to the development of basically new components, improvements in the design and manufacture of devices, lowering of the cost of devices and expansion of their uses will undoubtedly last for some years to come.

The improvement in the parameters of *diodes* is directly dependent on the electrical and physical characteristics of the semiconductor materials employed. Advances made in this field have allowed the creation of diversified devices ranging from high-power rectifiers rated at voltages of several kilovolts and currents of hundreds of amperes to microminiature pulse diodes with the reverse recovery time of tenths of a nanosecond.

The parameters of Zener diodes improve steadily. At present in wide use are high-stability temperature-stabilized precision Zener diodes which match conventional Weston cells in the accuracy of maintaining the reference voltage and even excell by far the latter in some other working characteristics.

On an increase is the production of diode mixers and varicaps, the cutoff frequency of which reaches several gigahertz.

In evidence is the tendency to further reduction in the cost of general-purpose diodes of various types, which is already reasonably low. This makes them practicable for use in most different fields of industry and agriculture.

Efforts are directed towards an increase in the working frequency and power of *transistors*. It is natural that the transistor working frequency can be increased by decreasing the base width, capacitance of *p-n* junctions and also the capacitance and inductance of the case. For small-power transistors this can be achieved through a reduction in the size of active regions. But for high-power transistor structures this approach is unsuitable because it does not help but rather hampers the solution of the problem of heat withdrawal and passage of heavy currents. This is why concurrent with a reduction in the emitter width one has to increase the emitter perimeter and complicate the geometry of the emitter region, making

it in the form of a "comb" or a net and, lastly, composing it of a large number of emitters. To avoid a nonuniform distribution of the current, use is made of thin-film resistors series-connected with each emitter. If the current injected by an emitter begins to grow, then, owing to the voltage drop across the series-connected resistor, the forward bias applied to the emitter-base junction near the given emitter will start to diminish, which in turn will bring down the current to a safe value. In the USSR and other countries small-power amplifying transistors are presently turned out with a cutoff frequency $f_{c\,max}$ of over 2000 MHz and a power of dissipation of up to one watt, and also switching transistors with a switching time in the order of 10 ns. It is anticipated that commercial transistors with $f_{c\,max} \geqslant 3$-5 GHz and with a correspondingly lowered noise figure from 5-6 dB at a frequency of 400 MHz (the noise level of present low-noise microwave transistors) to 1.5-3 dB at a frequency of 1-3 GHz will make their appearance in the near future.

The power of transistor amplifiers of many types intended for use in output stages of radio devices runs as high as tens and even hundreds of watts at frequencies in the range of tens and hundreds of megahertz. For example, in commercial production are now low-frequency silicon transistors with a dissipated power of 625 W and maximum collector current of 250 A, and silicon transistors with dissipated powers of 50 and 20 W at respective frequencies of 400 MHz and 1 GHz. In the next few years one can expect the appearance of transistors with powers of 100, 20 and 5 W at respective frequencies of 1, 2 and 5 GHz.

These parameters have become possible to attain thanks to certain advances in planar technology, in particular, the increase in the resolution of photomasking processes, ensuring the production of transistor structure components 2-3 μm wide.

The involved geometry of the above discussed transistors with record-breaking parameters and, besides, labour-consuming processes of their production make them unduly expensive. But then the cost of planar transistors with slightly lower power parameters attainable now due to the achievements in semiconductor technology and improvement of production equipment is quite reasonable. By now the prices of high-frequency planar transistors of many different types are comparable with the prices for alloy transistors.

Planar technology which embraces in principle a number of techniques enables one to process simultaneously hundreds and thousands of components on a single wafer, to make the production processes automatically controllable and to lower labour consumption in the fabrication of devices.

A new promising method of ion doping will soon be put to use in planar technology apart from the known method of alloying semiconductor materials by diffusion. The ion doping method

offers the ways of obtaining high-quality *p-n* junctions with a smaller depth of penetration and with a still smaller area than do the methods of photomasking and diffusion.

Extensive use has been made recently of the so-called diode or transistor *assemblies* (arrays) in which one crystal carries several similar-type devices. The desired interconnection of the diodes or transistors can be chosen by the operator, depending on the circuit diagram required.

These assemblies can also contain both *p-n-p* and *n-p-n* transistors isolated from one another either with additional back-biased *p-n* junctions or interlayers of dielectric, such as silicon dioxide.

In evidence is a fast growth in the field of *integrated circuits*. The production of integrated circuits was first started in 1962. Even the first results of their application surpassed the most optimistic expectations. The use of integrated circuits enables one to reduce drastically the overall dimensions and weight of radioelectronic apparatus, improve technical and service parameters and increase the reliability. These advantages of integrated circuits over discrete components have put them in the forefront and they have begun to rapidly supplant discrete semiconductor devices used in various apparatus.

The proportion of semiconductor integrated circuits in the total volume of semiconductor devices steadily increases and has now reached nearly one third of the total number. The share of MIS circuits grows appreciably too.

Particularly fast improvements are noticeable in the parameters of integrated circuits. One of the most important parameters of switching integrated circuits is the speed of response. At present there are integrated circuits intended for use in electronic computers with a signal propagation delay time of 3-5 ns, which gives grounds to expect the appearance of computers with a speed of 50-100 million operations in a second.

Further microminiaturization of radioelectronic devices calls for a reduction of the size of the integrated circuits proper and, correspondingly, the size of the active and passive components incorporated into them. This is why the level of integration is an important technical characteristic of any integrated circuit. In the first years of their appearance the integrated circuits comprised as a rule one or two tens of components in a single crystal measuring approximately 1.5 by 1.5 mm. As the semiconductor technology steadily develops the level of integration continues to grow and has already reached the second and third degree in the crystal of about the size mentioned above.

In designing such circuits it is necessary to solve the problem of removing the energy dissipated by a large number of components that occupy a little space in the crystal. The problem is tackled

both through the development of special structures and through a decrease in the consumption of power by the circuits, preserving at the same time the required operating characteristics.

The latter approach is especially important for the creation of not only ICs with a degree of integration of over 2 but also economical circuits employed in the devices whose consumption of power must be limited (space apparatus, for example). At the same time it is essential that the integrated circuits designed for operation in output stages should be made sufficiently powerful while their dimensions be kept small. Such conflicting demands attributed to the difference in the service conditions have resulted in the development of both micropower and high-power economical integrated-circuit amplifiers. At the present time the powers of various-type integrated circuits range from several milliwatts to tens of watts. It should be expected that this range will be further expanded.

Large-scale integrated circuits (LSIC) built up principally of MIS transistors have found particularly widespread use. Owing to a lower input power of MIS transistors, MIS-LSICs offer a higher level of integration in comparison with bipolar LSICs and at the same time ensure a speed of response sufficient for a wide range of applications.

MIS-LSICs built up of several thousand transistors in a crystal are now in commercial production. They represent in principle complete circuits of calculating, memory and other devices which have formed the basis for the creation of a new class of extremely small-sized microelectronic computers with improved parameters. Thus, key-actuated desk-top electronic computers and even "pocket" size computers containing only a few MIS-LSICs are now available.

The majority of integrated circuits are used as before in sophisticated machines, for example, in electronic computers. Advances in semiconductor technology and increase in the percentage yield have allowed an appreciable reduction in the cost of integrated circuits and thus opened the way for their use in comparatively small-capacity industrial and domestic apparatus.

If in the past five-year period the trend was towards gradual replacement of vacuum tubes with discrete semiconductor devices in most of domestic radioelectronic apparatus (television sets, for example), in the present five years the industry will undoubtedly shift to the production of integrated-circuit apparatus.

The creation of high-power switching *p-n-p-n* devices holds much promise. The best samples of commercially produced *thyristors* ensure the switching of currents of 300 to 400 A at a voltage of up to 1000 V. A major trend in the improvement of these devices will be towards an increase in the switching power and decrease in the switching time which at present comes to several microseconds. This will enable semiconductor devices to gain acceptance in new

fields of application, for example, high power supply, industrial automation, etc.

Silicon continues to be the most popular material for the production of semiconductor devices. The use of germanium has been sharply curtailed in the last few years. The share of germanium transistors and diodes out of the total number of devices produced was reduced by 70—80 per cent and does not exceed 20 per cent of the total volume. The percentage of silicon devices steadily grows.

The development and use of novel semiconductor materials open up basically new possibilities for semiconductor electronics.

Gallium arsenide offers the greatest promise as a new material. Commercial production methods have been recently elaborated which made it possible to fabricate quite suitable monocrystals and to grow doped epitaxial layers of gallium arsenide.

Although gallium arsenide is inferior to silicon in its technological potentialities, it is indispensable in the fabrication of a number of devices, such as Gunn effect oscillators. This is attributable to the peculiarities of its energy-band structure and also to a higher mobility of electrons and greater width of the band gap compared to those in germanium and silicon.

Gallium arsenide diode structures become especially popular. Varactors with a maximum frequency of 150 GHz, diode mixers with a working frequency of 10 GHz and tunnel diodes with a maximum-to-minimum current ratio of over 10 are produced with the use of gallium arsenide as a basic material.

Gallium arsenide microwave diodes successfully compete with microwave diodes made of silicon, while gallium arsenide light-emitting diodes have ousted silicon light-emitting diodes to the second place and now hold the lead. Diodes employed in miniature devices emit the energy of 15 mW within a sector of 2.5° at a diode current of 50 mA. High-power light-emitting diodes used in communication systems and for lighting have a luminous efficiency in the order of 200 mW/A.

There is no doubt that the power of laser diodes will be substantially increased and their field of application broadened. The maximum pulse power of gallium arsenide lasers now reaches tens and hundreds of watts.

Only 13 years have passed since the Gunn effect was discovered. But the series production of Gunn diodes is now under way in many countries. Their power is as great as 0.5 W at a frequency of up to 10 GHz and a pulse power of 1 MW at a frequency of 1 GHz. It should be expected that in the near future the working frequencies and powers will undoubtedly grow by approximately one order of magnitude.

The results are less conspicuous when use is made of gallium arsenide in the fabrication of bipolar transistors, though it was

anticipated in early stages of its employment that these transistors would excell germanium and silicon transistors in frequency and temperature characteristics. The results of experiments do not promise so far much success in this field. The reproducibility of the characteristics of gallium arsenide transistors has been found to be poor and the frequency properties proved far from those predicted from theoretical calculations. The inadequate results can be apparently accounted for by the presence of deep carrier traps in gallium arsenide. The ways of controlling the behaviour of these traps have not yet been found.

Investigations conducted in the last few years have shown that the use of gallium arsenide for the production of MIS transistors is more promising.

Gallium arsenide enhancement-mode type transistors encompass a frequency range from 4 to 20 GHz and are not inferior to bipolar devices in amplification and noise factors. Thus, the noise factor of laboratory samples of gallium arsenide MIS transistors at a frequency of 4 GHz is 2-3 dB (3.8 dB for bipolar transistors), at a frequency of 8 GHz, 4dB (over 6 dB for bipolar transistors), and at a frequency of 12 GHz, 5 dB (bipolar transistors do not amplify at this frequency).

Extended research is in progress for the creation of gallium arsenide integrated circuits and functional devices.

Among the above discussed gallium arsenide devices, only varicaps, microwave diodes, light-emitting and laser diodes are manufactured in production quantities, but there is no doubt that the list will be extended significantly in the near future.

Intensive research is under way on the binary compounds of the elements of groups II and VI (CdS, PbS, ZnS, CdSe), groups III and V (GaAs, GaP, InSb, etc.), groups III and VI (In_2Te_3, Ga_2Te_3, etc.), trinary systems of groups III and V (GaAlAs, $GaAsP_2$, etc.), silicon carbide, and vitreous and many other semiconductor compositions. They are used as basic materials for constructing new interesting devices and improving the parameters of known devices.

The search for the creation of the devices operating on new principles is being continued. It should be pointed out in this connection that in 1972 ultrahigh frequency devices were developed on the basis of the Josephson effect discovered in 1961 which manifests itself as tunneling of fixed electrons at ultralow temperatures. Such devices have a junction formed between a pair of superconductor contact plates, as niobium ones, which are interspaced with a thin dielectric or a semiconductor layer. The generated frequency depends on the voltage applied to the device (20 mV for 10 GHz and merely 0.2 mV for 100 GHz) and can reach 700 GHz. Josephson-junction devices may be utilized for switching purposes. The swit-

ching speed may equal several picoseconds and the input power only several microwatts. The main disadvantage lies in the necessity of maintaining an ultralow temperature at about 4° K.

Let us dwell in somewhat greater detail on the research work at evolving *image storage devices.* Possibilities are investigated of utilizing the phenomena of near-breakdown electroluminescence observed in the semiconductor crystalline phosphors, and injection electroluminescence which occurs in *p-n* junctions on passing the current in the forward direction.

Image storage devices operating on the first principle represent a solid panel with several layers, including a control ZnO layer negatively charged in advance. On projecting an image, individual portions of the layer discharge and leave on the ZnO layer surface the stored charge image controlling the external layer of the luminophor. The image can be preserved for an hour or erased by applying a large negative potential. Laboratory models of these evices measure up to 30 × 30 cm^2 and have a resolution of 400 to 800 lines.

The devices based on after-injection electroluminescence represent matrices built up of a sufficient number of light-emitting diodes and serve mainly for visual letter-digit indication. Matrices are now available that comprise 35 diodes (5 × 7) and are capable of reproducing any letters and figures. The diameter of each luminous point is 0.5 mm with a distance between points of 1 mm.

Capacitive coupling devices (CCD) also gain recognition in the creation of image formation systems. In the simplest variant, they consist of a chain of MIS capacitors arranged on a silicon substrate. On illuminating a MIS capacitor, the latter generates a charge whose storage and transfer are effected by applying appropriate voltages to the capacitor electrodes. CCD matrices contain up to 10,000 elements, which permits obtaining sufficiently good images in phototelegraphy. CCDs are also applied in the construction of memory devices.

In devising an electrically programmed permanent memory, use is made of one more new element—a floating-gate MIS transistor. The transistor gate has no external terminal and is coated at the top with a dielectric layer. Information is recorded by applying to the transistor drain a voltage of 30 to 50 V which is sufficient to initiate an avalanche breakdown near the drain. Hot electrons are injected from the breakdown region across a dielectric layer 1000 Å thick to be entrapped by the gate. The charge of these electrons induces a channel between the source and the drain. Since the gate is isolated from the other electrodes, the charge on the gate and the channel can be stored for years. The information is cancelled by X-ray irradiation.

The devices described above open up a completely new field of application of semiconductor apparatus and have a great future.

As has been mentioned earlier, to improve further the frequency properties of devices as well as increase the level of integration of integrated circuits it is necessary to reduce the linear dimensions of active regions of semiconductor devices. But the reduction of component size to less than 0.5-1 μm does not seem feasible through the use of conventional methods because these dimensional values become commensurable to the value of the wavelength of the light employed in photolithography.

The problem of overcoming this barrier is being tackled with the methods of electronic photolithography, that consist in drawing of a pattern on the photoresist by an electron beam and direct introduction of impurities into a semiconductor by an ion beam. The technique of ion doping is the first step on the way to the development of semiconductor devices on the basis of literally *molecular electronics*.

If we take into consideration steady advances in mechanical engineering, chemistry, optics and other industries, offering considerable scope for the creation of more elaborate automatic equipment for the production of semiconductor devices, there can be hardly any doubt that already in the immediate future we shall witness a still more vigorous progress in the development of semiconductor industry and more extensive use of semiconductor devices.

APPENDIX 1

Fundamental Constants

Planck's constant

$h = 6.62 \times 10^{-27}$ Erg·s $= 6.62 \times 10^{-34}$ J·s $= 4.14 \times 10^{-15}$eV·s

Boltzmann's constant

$k = 1.38 \times 10^{-16}$ Erg/deg $= 1.38 \times 10^{-23}$ J/deg $= 0.86 \times$
$\times 10^{-4}$ eV/deg

Mass of electron

$m_0 = 9 \times 106 \times 10^{-28}$ g

Electron charge

$q = 1.602 \times 10^{-19}$ C

Electron diameter

$D = 3.8 \times 10^{-13}$ cm $= 3.8 \times 10^{-15}$ m

Permittivity of free space

$\varepsilon_0 = 8.86 \times 10^{-14}$ F/cm $= 8.86 \times 10^{-12}$ F/m

Permeability of free space

$\mu_0 = 1.257 \times 10^{-8}$ H/cm

Speed of light

$c = 2.99778 \times 10^{10}$ cm/s

Avogadro number

$N = 6.02 \times 10^{23}$ molecule/mole

1 eV $= 1.602 \times 10^{12}$ Erg $= 1.602 \times 10^{-19}$ J

APPENDIX 2

Properties of Semiconductors at 300°K

Parameter	Ge	Si	GaAs	InSb	SiC
Density, g/cm^3	5.32	2.33	5.3	5.78	3.17
Atomic weight	72.6	28.08	72.5	118.0	20.0
Density of atoms, cm^{-3}	4.4×10^{22}	5.0×10^{22}	1.3×10^{22}	1.4×10^{22}	4.7×10^{22}
Lattice constant, Å	5.66	5.43	5.63	6.48	4.36
Melting point, °C	937	1417	1238	536	2700
Relative permittivity	16.3	11.7	11.1	16.0	7.0
Band gap width, eV	0.67	1.11	1.43	0.18	3.1
Intrinsic carrier density, cm^{-3}	2.4×10^{13}	1.45×10^{10}	$9\cdot10^{14}$	—	—
Electron mobility, cm^2/V·s	3800	1300	8500	7700	150

Continued

Parameter	Ge	Si	GaAs	InSb	SiC
Hole mobility, $cm^2/V \cdot s$	1800	470	440	700	30
Electron diffusion factor, cm^2/s	98	33.6	220	2000	2.6-3.9
Hole diffusion factor, cm^2/s	47	12.3	11.2	18	0.5-0.8
Linear expansion coefficient, 1/°C	5.8×10^{-6}	2.5×10^{-6}	5.9×10^{-6}	—	—

APPENDIX 3

Thermo-emf of Semiconductors

Substance	α, μV/deg	Substance	α, μV/deg
Tl_2S	—780	FeS	+26
MoS	—770	CdO (p)	+30
WO_3	—740	Bi_2S_3	+80
CuO_3 (n)	—700	Sb_2Te_3	+100
Fe_2O_3	—613	Bi-Sb-Te	+160
FeO	—500	Bi_2Te_3	+170
SiC (n) green	—300	Bi_2Se_3	+200
FeS_2	—200	SbZn	+200
$CoSb_3$	—200	NiO	+240
Bi-Te-Se	—170	MoS_2	+300
PbS	—160	Mn_2O_3	+385
PbTe	—160	CoO	+450
SnO	—140	SiC (p) black	+800
CdO (n)	—40	Se	+1000
CuS	—7	CuO (p)	+1120
		Cu_2O	+1200

APPENDIX 4

Thermo-emf of Metals and Alloys

Material	α, μV/deg	Material	α, μV/deg
Bismuth	—68.0	Tin	—0.2
Constantan	—38.0	Magnesium	0
Copel	—38.0	Lead	0
Nickel	—20	Silver	+2.7
Nichrome	—18.0	Gold	+2.9
Alumel	—17.3	Zinc	+3.1
Palladium	—8.9	Tungsten	+3.6
Sodium	—6.5	Molybdenum	+7.6
Platinum	—4.4	Iron	+15.0
Mercury	—4.4	Chromel	+24
Aluminium	—0.4	Antimony	+43

APPENDIX 5

Relationships among SI and cgs Units

Quality	Unit of measurement		SI-to-cgs unit ratio
	SI	cgs	
Length	m	cm	10^2
Mass	kg	g	10^3
Time	s	s	1
Force	N	dyne	10^5
Pressure	N/m^2	$dyne/cm^2$	10
Energy	J	Erg	10^7
Current	A	—	3×10^9
Quantity of electricity (electric charge)	C	—	3×10^9
Potential difference	V	—	1/300
Power	W	Erg/s	10^7
Electric field strength	V/m	—	$1/(3 \times 10^4)$
Electric field induction	C/m^2	—	3×10^5
Resistance	ohm	—	$1/(9 \times 10^{11})$
Capacitance	F	—	9×10^{11}
Inductance	H	cm	10^9
Magnetic field strength	A/m	Oe (oersted)	$4/10^3$
Magnetic induction	T	G (gauss)	10^4
Flux of induction	Wb	M (maxwell)	10^8

APPENDIX 6

Standard Graphical Symbols for Semiconductor Devices (according to State Standard, ГОСТ 2730-68)

1		Semiconductor diode, rectifier. Vertex of the triangle points the direction of maximum conduction
2		Tunnel diode
3		Backward diode. Arrow indicates the direction of maximum conduction
4		Zener (avalanche) diode
5		Varicap

6		Thyristor with a lead brought out from an *n*-region
		Thyristor with a lead brought out from a *p*-region
7		Four-layer diode, *p-n-p-n* switch
8		*n*-base unijunction transistor
9		*p*-base unijunction transistor
10		*p-n-p* triode
11		*n-p-n* triode
12		*p*-type thyristor, *p-n-p-n* switch
13		*n*-type thyristor, *n-p-n-p* switch
14		*n*-base channel-type transistor

15		*p*-base channel-type transistor
16		Insulated-gate channel-type transistor with a flat junction coupling the source and the drain with an *n*-base (enhancement-mode type)
17		Insulated-gate channel-type transistor with a flat junction coupling the source and the drain with a *p*-base (enhancement-mode type)
18		Insulated-gate channel-type transistor with a flat junction coupling the source and the drain with an *n*-base (depletion-mode type)
19		Insulated-gate channel-type transistor with a flat junction between the source and the drain and with an ohmic lead brought out from a *p*-base (depletion-mode type)
20		Photoresistor
21		Photodiode
22		Semiconductor photocell
23		*p-n-p* phototriode

24		*n-p-n* phototriode
25		Galvanomagnetic semiconductor device. Hall-effect device

REFERENCES

1. Bazarov V. K. *Poluprovodnikovye lazery i ikh primenenie* (Semiconductor Lasers and Their Applications). "Energiya" Publishers, 1969.
2. Bube, Richard H. *Photoconductivity of Solids*. N.Y., Ind. 1960.
3. Efimov, I. V., Kalman, I. G., Martynov, V. I. *Nadezhnost integralnykh skhem* (Reliability of Integrated Semiconductor Circuits). Komitet standartov Publishing House, 1969.
4. Fedotov, Ya. A. *Osnovy fiziki poluprovodnikovykh priborov* (Fundamentals of the Physics of Devices). "Sovetskoe radio" Publishers, 1969.
5. Gavrilov, R. A., Skvortsov, A. M. *Technologiya proizvodstva poluprovodnikovykh priborov* (Technology of Production of Semiconductor Devices). "Energiya" Publishers, 1968.
6. Gavr lov, R. A., Skvortsov, A. M. *Osnovy fiziki poluprovodnikov* (Fundamentals of Semiconductor Physics). "Mashinostroenie" Publishers, 1966.
7. Gorodetsky, A. F., Kravchenko, A. F. *Poluprovodnikovye pribory* (Semiconductor Devices). "Vysshaya shkola" Publishers, 1967.
8. Kolosov, A. A., Gorbunov, Yu. I., Naumov, Yu. E. *Poluprovodnikovye tverdye skhemy* (Semiconductor Solid Circuits). "Sovetskoe radio" Publishers, 1965.
9. Krasilov, A. V., Trutko, A. F. *Metody rascheta tranzistorov* (Transistor Design). "Energiya" Publishers, 1964.
10. Kurnosov, A. I., Yudin, V. V. *Tekhnologiya proizvodstva poluprovodnikovykh priborov* (Technology of Production of Semiconductor Devices). "Sudostroenie" Publishers, 1965.
11. Middlebroock, R. D. *An Introduction to Junction Transistor Theory*. N.Y., Ind. 1957.
12. Nikolaevsky, I. F. *Ekspluatatsionnye parametry i osobennosti primeneniya tranzistorov* (Service Parameters and Uses of Transistors). "Sovetskoe radio" Publishers, 1963.
13. Pasynkov, V. V., Chirkin, L. K., Shinkov, A. D. *Poluprovodnikovye pribory* (Semiconductor Devices). "Vysshaya Shkola" Publishers, 1966.
14. Pikus, G. E. *Osnovy teorii poluprovodnikovykh priborov* (Fundamentals of the Theory of Semiconductor Devices). "Nauka" Publishers, 1965.
15. Ryvkin, S. M. *Fotoelektricheskie yavleniya v poluprovodnikakh* (Photoelectric Phenomena in Semiconductors). "Fizmatgiz" Publishers, 1963.
16. Stepanenko, I. P. *Osnovy teorii tranzistorov i tranzistornykh skhem* (Fundamentals of the Theory of Transistors and Transistor Circuits). "Energiya" Publishers, 1973.
17. Valiev, K. A., Karmazinsky, A. N., Korolev, M. A. *Tsyfrovye integralnye skhemy na MDP tranzistorakh* (Digital MIS Transistor Integrated Circuits). "Sovetskoe radio" Publishers, 1971.
18. Vaysburd, F. I. *Poluprovodnikovye pribory* (Semiconductor Devices). "Svyaz" Publishers, 1966.
19. *Thin Film Microelectronics. The Preparation and Properties of Components and Circuit Arrays*. Edited by L. Holland. F. Inst. P. Head of the Central Research Laboratory. London, 1965.
0. *Plenochnaya microelectronika* (Film Microelectronics). Collection of articles edited by Lukin, F.V., ed. 1. "Sovetskoe radio" Publishers, 1967.
21. *Poluprovodnikovye diody* (Semiconductor Diodes). Collected works, edited by Goryunov, N.N., Nosov, Yu. P. "Sovetskoe radio" Publishers, 1968.
22. *Tranzistory* (Transistors). Reference book, edited by Nikolaevsky, N. F. "Svyaz" Publishers, 1969.

Accumulation
carrier, 78
Accumulation region
charge carrier, 58
Admittance
forward transfer, 123
input, 123
output, 124, 125
representation of, 123
reverse transfer, 123
Amplification
current, 127
factor of, 178
Amplifier
integrated circuit, 251
pulse, 142, 143
transistor, 146, 249
Anodization method, 242
Avalanche multiplication, 70, 114, 115
carrier, 53, 134
factor of, 115

Band gap width, 19-22, 70, 142, 144, 203
Barrier layer width, 25
Binistor (see *p-n-p-n* Switch), 180-185
Boundary
base-emitter, 113
between *p*-semiconductors, 22
Boundary region, 25
Breakdown
avalanche, 51-54, 60, 82-86
collector junction, 114,134
p-i-n junction, 101
p-n junction, 50, 82
thermal, 51-54, 65, 71, 83
tunnel (Zener), 51-54, 60, 82 85, 86

Capacitance
charge (barrier-layer), 54, 55, 74, 87, 90
collector, 150, 159, 166
diffusion, 55, 56, 87, 116, 117
diode, 67, 68, 74, 88
emitter, 166
emitter diffusion, 121
gate, 175
gate-drain parasitic, 175
interelectrode, 171
p-n junction, 54, 65
transistor, 116
tunnel diode, 95
Capacitance ratio, 88
Capacitive reactance, 56, 88
Capacitor
film, 241
MIS, 90 102
voltage-variable, 87, 90
Capture cross section
electron, 205
Carrier generation, 45, 46, 204, 206
Carrier mobility, 27-31, 59, 99, 163 198, 199
Carrier multiplication, 181
Cathode evaporation, 241
Cathode sputtering, 242
Cathodoluminescence, 221
Centre
acceptor, 57, 59
donor, 57, 58
recombination, 205
trap, 120, 205
Charge distribution
in *p-n* junction, 35, 36
Coefficient
absorption, 204, 209
conductivity, 190
linear expansion, 72
Peltier, 188
Seebeck, 188
surface recombination, 108
thermoelectric, 186, 189
Thomson, 189
Conductance
channel, 173, 174
Conduction
electron, 33
heat, 190
hole, 33
intrinsic, 37
Conduction band, 19-22, 58, 85, 97
Conductivity
dark, 203, 206
light, 203
negative differential, 61
surface, 32

Connection, transistor
common-base, 126-133, 139
common-collector, 126-133, 139
common-drain, 177, 178
common-emitter, 126-133, 139
common-gate, 177, 178
common-source, 177, 178
Contact
between two *n*-semiconductors, 22
between two semiconductors, 20, 24
metal-semiconductor, 27
ohmic, 25, 56
semiconductor-metal, 22-26, 31
Contact phenomenon, 19
Continuity equation, 43
Covalent bond, 57
Current
collector, 114, 135, 138, 144, 150
conduction, 27
diffusion, 27, 41-44, 60, 92, 94
drain, 170, 174
electron drift, 40
emitter, 150
forward, 27, 29, 41, 50, 65-70, 81, 94
fourpole, 122
generation, 46
hole drift, 40
leakage, 66
peak, 94-97
rectified, 67, 68
regulation, 83-86
reverse, 29, 31, 46, 49, 50, 53, 66-70, 81
saturation, 114, 174
surface recombination, 108
thermal, 30, 45, 46, 49, 55, 66
tunneling, 93, 94
valley (excess), 94-97
Current gain, 127, 136, 162
alpha, 105, 109, 115, 120
calculation of, 105
frequency dependence of, 117
Current pulse shape
in Gunn diode, 98
Current transfer ratio
for transistor, 127, 141, 165
Current-voltage characteristic
backward diode, 95
calculation of, 43-46
dinistor, 181, 182
diode, 69, 71, 76, 83-86, 92
field-effect transistor, 169, 170
forward, 47, 48, 60, 76, 83, 92
Gunn diode, 101
photocell, 217
photodiode, 214
p-n junction, 43, 51, 53
reverse, 51, 54, 82, 83
semiconductor-metal contact, 29-31
thyristor, 182, 183
tunnel diode, 96
unijunction transistor, 179, 180
Cutoff frequency
alloy transistor, 150, 153
amplifying transistor, 249
current-gain, 118, 127, 136, 146, 150, 153, 165
diffused transistor, 153
planar-epitaxial transistor, 162

Debye length, 172
Density
acceptor, 162, 163,
carrier (hole, electron), 19-58, 91, 100, 108, 113, 142
collector current, 108
conduction current, 27
diffusion current, 27
drain current, 174
equilibrium, 28, 29, 33, 41-43, 47, 78, 203, 204
excess carrier, 206
free carrier, 198, 203
impurity, 23, 25, 36
intrinsic, charge carrier, 203
mobile carrier charge, 174
nonequilibrium, 203, 204, 207
recombination centre, 207
reverse current, 82
space charge, 24, 25, 39
of stages, 49, 57
surface donor impurity, 162, 166
Density distribution
carrier, in *p-n-p* transistor, 106
hole, in base region, 112
impurity, laws of, 62-62
Density gradient
acceptor, 39
charge carrier, 145
donor, 39
electron, 40
hole, 40, 113
impurity, 153, 165
steady-state, hole, 112

Depletion layer width, 25, 31, 43, 111
Deposition
sequential, 234
Diffusion
charge carrier, 33
outward, 159, 160
process of, 153-157
selective, 154, 234, 176
sequential, 171
sequential double, 145
Diffusion constant
electron, 163
hole, 164
impurity, 63
Diffusion factor
electron, 28, 43
hole, 43
minority carrier, 136
Diffusion length
electron, 43
hole, 42, 43
majority carrier, 136
minority carrier, 56, 103, 136
Dinistor (See *p-n-p-n* Switch), 180-185
Diode
alloyed-diffused, 64, 82
avalanche (Zener), 61, 82-87
avalanche transit time, 61
diffused-junction, 62, 82
double-base junction, 179, 180
flat (junction), 61, 62
fused-junction, 62
Gunn-effect, 61, 97, 252
IMPAAT, 97, 101, 102
injection-type, 224
light-emitting, 221-223
micro-alloy junction, 65, 66, 82
point-contact, 61, 64-66, 82
point-junction, 61, 64, 73, 74, 77
pulse (switching), 61, 78-82
rectifier, 61, 75, 70
surface-barrier (junction), 61, 66
theory of, 30
tunnel, 16, 61, 92, 97
voltage-variable capacitance, 61, 87-89
Domain shape
in Gunn diode, 99
Double layer
at contact, 24

Effective mass
charge carrier, 187
electron, 29

Einstein relation, 46, 163
Electroluminescence, 221, 222, 254
injected, 223
Electroluminescent cell
structure of, 223
Electron affinity, 24
Electron charge, 30
Electron free path, 31
Electron-hole pair, 14, 51, 52, 101, 203
generation of, 204
Electron kinetic energy, 29
Electron pumping, 224
Electron velocity
mean, 29
Electrostatic potential, 21, 35, 38
Emission intensity
light-emitting diode, 224
Emitter injection efficiency, 107, 110, 111, 120
Energy band bending, 20, 21
Energy-band diagram of
acceptor states, 58
contact, 20-24, 35, 66
ohmic contact, 56
semiconductor-metal contact, 189
tunnel breakdown, 51
tunnel diode, 82
tunneling junction, 51
Energy barrier, 21, 22
Energy level, 19, 20, 58
inverse population of, 224
Epitaxial growing, 159, 160
sequential, 171
Epitaxial layer, 162
Epitaxy
liquid-phase, 246
vapour-phase, 246
Equilibrium distribution
charge carrier density, 34
space charge density, 34
Equivalent circuit
current generator, 129
diode, 66-68
fourpole, 124, 125
parameters of, 130
photocell, 218
T-type, transistor, 109, 110, 128, 129, 131
varicap, 88
voltage generator, 129
Evaporator
electron bombardment, 239
plasma, 239

Failure rate, 228, 229
component, 233

Fermi level, 19, 22, 30, 32, 93
Fermi potential, 20, 21, 34, 35
Fick equation, 62
Field strength
    in diode, 100
    in domain, 99
    at *p-n* junction, 50-52
Fourpole
    equations of, 122-126, 130
    representation of, 122-126
Frequency response
    photodiode, 216
    transistor, 146

Hall constant, 197
Hall effect, 197, 199
Hall emf, 197, 200
Hall generator, 199-201
    parameters of, 200
    response time of, 200-201
Heat transfer, 65
h Parameters, 125

Image storage device, 254
Impedance
    backward, 67, 68
    dimension of, 123, 124
    forward transfer, 124
    input, 124, 125, 128, 177
    output, 124, 128, 177
    representation of, 124
    reverse transfer, 124
Impurity
    acceptor, 19, 33
    donor, 19, 33
Impurity distribution in
    alloy transistor, 145
    diffused-alloy transistor, 154
    drift transistor, 146
Impurity level, 22
Injection level, 46, 47
Injection ratio, 46
Integrated circuit, 250, 251
    bipolar, 231, 233
    density of, 231, 232
    epitaxial film, 245
    film, 234, 235-241
    hybrid, 235
    integration degree (level) of, 232, 237
    large-scale (LSIC), 232, 251
    MIS, 231, 233, 237
    MIS-LSIC, 251
    monolithic, 234
    packaging, 232
    technology of, 231, 234

Ion doping, 249, 255
Ion implantation, 176
Ionization
    collision (impact), 51, 52, 114, 224
Ionization energy
    of impurities, 33

Josephson effect, 253
Joule effect, 190
Joule heat, 194, 195
Junction
    abrupt, 37-41, 60, 64, 116
    alloyed, 77
    asymmetrical, 36-39, 47, 59
    breakdown of, 50-53
    capacitance of, 54
    collector, 103, 104, 134, 142
    diffused, 77
    emitter, 103, 104, 114, 115, 142
    emitter-base, 104
    flat, 65
    forward-biased, 45, 85
    graded, 37, 39, 55, 60
    reverse-biased, 51, 85, 168, 250
    reverse-biased collector, 114, 115
    symmetrical, 36-38
    thermoelement, 194
    tunneling, 51

Laser, 16, 252
    injection, 226
Lattice constant, 33
Lattice site, 57
Lifetime
    electron, 43, 46, 55, 80, 81, 163, 204-205
    hole, 43, 46, 55, 106, 163, 204, 205
    minority carrier, 139
Light intensity
    incident, 204
Lorentz force, 197
Luminescence, 221
    effect of, 221
    recombination, 223

Magnetic induction, 197, 200
Mobility ratio, 47
Mode of operation
    active, 137
    cutoff, 137-139
    depletion, 173

enrichment, 173
LSA, 100
saturation, 135, 139, 182
small-signal, 122
Multiplication factor, 53

Noise
intrinsic, 61, 91, 119
leakage, 120
low-frequency, 120
partition, 119
shot, 119
thermal, 119
transistor, 119, 120
Noise bandwidth, 86
Noise factor, 97, 119
Noise level
MIS transistor, 178, 179,
Noise power, 119

Operating (quiescent) point
transistor, 137, 138, 143
Oscillation period, 100, 101
Oscillator
diode, 96
Gunn, 100, 101, 252
transistor, 146

Packing density, 177
circuit, 229-231
Peltier effect, 187, 188
Peltier heat, 194, 195
Penetration depth, 171, 250
atom, 62
impurity, 148
ion, 176
Photocell, 203, 213, 218
efficiency of, 219
open-circuit condition of, 217
short-circuit condition of, 217, 219
spectral response of, 218
Photoconduction, 203, 205
Photoconductivity
fall time of, 207, 208
infrared decay of, 208
intrinsic, 210
rise time of, 207, 208
semiconductor, 204
spectral distribution of, 209
steady, 204, 206
Photocurrent
dark, 207, 208
light, 207, 208, 211
Photodiode, 203, 213, 214, 216
frequency response of, 216
parameters of, 215
spectral sensitivity of, 216
threshold sensitivity of, 215
Photoeffect
extrinsic, 203
intrinsic, 203
Photoelectric emission, 203
Photoelectromotive force, 15
Photoluminescence, 221
Photomasking, 155
Photoresist, 155, 157, 177
Photoresistor, 210
dark resistance of, 212
design configuration of, 211
integral specific sensitivity of, 212
light resistance of, 212
parameters of, 213
relaxation time of, 213
Photosensitivity, 210
semiconductor, 203
Phototransistor (phototriode)
collector dark current of, 219
sensitivity of, 220
Piezoelectric delay line, 233
Piezoelectric transducer, 233
Pisarenko formula, 187
Poisson equation, 24, 37, 39
p-n-p-n Switch, 168
four terminal (binistor), 180
parameters of, 184
three terminal (thyristor), 180
two-terminal (dinistor), 180
Potential barrier, 21, 24, 25, 27, 31, 34, 41, 43
contact, 35
height of, 26, 29, 30, 41, 59
intrinsic, 23
p-n junction, 181
Potential distribution, 41, 43, 99
diagram of, 20, 21
equilibrium, 41
Power dissipation, 51, 69
diode, 72
transistor, 141, 142, 146, 249
Power gain, 104
Proportionality factor, 201
Pulse response
transistor, 136
Pulse waveform
transistor, 140
Punch-through
p-n junction, 115

Quality factor of
varicap, 89
Zener diode, 84, 85

Quantum transition
electron, 221
nonradiative, 221, 222
radiative, 221, 222
Quantum yield, 204

Radiation
coherent, 224
electromagnetic, 221
emitted, 222
luminescent, 221
recombination, 222-224
Radiant energy transfer, 203
Recombination
bulk, 108
carrier, 45, 80
linear, 205-208
probability of, 205
quadratic, 207, 208
radiative, 224
surface, 108
Recombination emission, 222
Recombination loss, 134
Recombination rate, 71, 204-207
Rectification at
metal-semiconductor contact, 27
p-n junction, 40
semiconductor-metal contact, 23-26
Rectification ratio, 31, 50, 56, 75, 243
Rectifier
diode, 70-75
polycrystalline, 75-77
Relative permittivity, 24
gate insulator, 174, 175
Relaxation constant
for linear recombination, 207
Reliability
semiconductor device, 226-229
Resistance
barrier, 26, 77
base, 65, 78, 109, 111, 114, 150
bulk, 47, 66, 112
collector, 109-111, 134
differential, 170
differential dynamic, 84, 97
differential negative, 53, 54, 82, 95, 98, 180
diffusion, 111
emitter, 109, 110, 134, 136
forward, 55, 78
input, 171
intrinsic, 180, 181
leakage, 67
load, 81, 86
ohmic, 111, 112
pulse, 78
reverse, 66, 74, 88
saturation, 159, 162
space charge layer, 26
spread, 66, 77
static, 84
thermal, 53, 142
Resistivity
base, 52, 164
diode, 70, 74
semiconductor, 52, 53
surface, 235
transistor channel, 171

Schottky diffusion theory, 27, 31
Seebeck effect, 186
Semiconductor
degenerate, 93, 95
extrinsic, 19
high-resistivity, 39
intrinsic, 19, 49
low-resistivity, 32, 39
n-type, 19, 25, 33, 59
p-type, 25, 33, 59
Semiconductor light source, 221-226
Solubility limit
diffusate, 62
Space charge
light-generated, 220
negative, 21
positive, 21, 24
Space charge layer (region), 168, 169
injected carrier, 46-48
negative, 99
width of, 23-40, 59, 61, 71, 101, 102, 112
Spectral sensitivity
photoresistor, 210
Speed of response
hybrid *IC*, 235
light-emitting diode, 223
MIS transistor, 175, 176
planar-epitaxial transistor, 162
tunnel diode, 94
varicap, 102
Stabistor, 83, 86
Surface state
acceptor, 58
density of, 57, 58
property of, 57

Temperature coefficient of
current, 76
resistance, 241
voltage, 84, 85, 102
Thermal current
temperature dependence of, 49

Thermal emf, 190, 202
generation of, 186
Thermal equilibrium, 191, 195
Thermal generation of carriers, 203
Thermal potential, 28
Thermoelectric cooler, 152, 188, 194
Thermoelectric generator, 190-202
Thermoelectric phenomenon, 186
Thermoelectric trap, 196
Thermoelectricity, 186
Thermoelement, 190-196
efficiency of, 193
Thick-film component, 243
Thin-film component, 243
Thomson effect, 189
Thyristor (see *p-n-p-n* switch), 180-185, 251
Time
dielectric relaxation, 91, 100, 201
domain buildup, 100
domain transit, 101
drift, 91
electron tunneling, 94
fall (decay), 139, 206, 207
forward transient, 78-82
response, 200
reverse recovery, 78-82, 102
rise (turn-on), 139, 296, 207
storage, 82, 139
transit, 146, 165
trapping, 120
Time characteristic
dinistor, 182
transistor switch, 139
Time constant
carrier-diffusion, 138
feedback, 161
Transadmittance, 171, 174
Transfer
carrier, 42
electron, 24, 27, 40
forward, 123
reverse, 123
Transfer ratio
forward current, 125
reverse voltage, 125
Transistor
alloy, 145-150, 153, 157, 167
bipolar, 178
channel-type, 168, 173
current of, 114
diffused-alloy, 153
diffusion, 144, 147, 153
drain current of, 171
drift, 144-146, 153, 166, 167
field effect, 168-179
film insulated-gate field-effect, 244
floating-gate MIS, 254
forward transfer characteristic of, 131
fused, 118
gallium arsenide, 144
germanium, 144, 148, 150, 161
germanium alloy (fused) *p-n-p*, 144
germanium *p-n* junction field effect, 171
indium phosphide, 144
induced channel (MIS), 172, 173
input characteristic of, 131, 135
input resistance of, 171
insulated-gate field-effect (MIS), 171-173, 177
interelectrode capacitance of, 171
intrinsic resistance of, 109
junction, 112
low-noise microwave, 249
maximum current of, 141
mesa, 154, 157, 162
microalloy, 152
MIS, 59
MIS, *n*-channel, 172, 173
MIS, *p*-channel, 172, 173
planar, 154-159, 162, 167, 249
planar epitaxial, 162, 167
planar, parameters of, 161, 162
point-contact, 16
pulse response of, 132
reverse transfer characteristic of, 131
saturation coefficient of, 138
silicon, 144, 148, 150, 157, 161, 249
silicon field-effect, 175
static characteristics of, 131, 133
T equivalent circuit of, 128-131
unijunction, 168, 179
unipolar (see Field-effect)
voltage of, 114
Transistor operation region
active, 137, 143
avalanche multiplication, 137
cutoff, 137
saturation, 137, 138
Transistor switch, 136, 137
Transport factor (efficiency), 107, 108, 110, 120
Tunnel effect, 51

Upper frequency limit
in diode, 69

Vacuum deposition, 241
Vacuum evaporation, 223, 240
Valence band, 19-22, 58, 85, 93
Varicap, 61, 87-90
Velocity
drift, minority carrier, 118
electron, 29, 98
electron thermal, 205
mean carrier, 198
minority carrier, 118
surface recombination, 108, 109, 136
Voltage
bias, 43
breakdown, 50-53, 69, 70, 84-86
collector, 105, 112, 141, 150
collector breakdown, 115
cutoff, 169
drain, 169, 170
emitter-collector breakdown, 115
formation, 65
gate, 169
Hall, 201
mean-square noise, 119
peak, 94-97
pump, 91
residual, 182
saturation, 138, 139
threshold, 173, 177
valley, 94-97
Voltage gain, 178

Work function, 22-25, 30, 203

y Parameter, 123

z Parameter, 124

TO THE READER

Mir Publishers welcome your comments on the content, translation, and design of the book.

We would also be pleased to receive any suggestions you care to make about our future publications.

Our address is:
USSR, 129820,
Moscow, I-110, GSP,
Pervy Rizhsky Pereulok, 2,
Mir Publishers

Printed in the Union of Soviet Socialist Republics

OTHER BOOKS FOR YOUR LIBRARY

Electronic Devices

V. DULIN, Cand. Sc.

425 pp. 14 X 21. 5 cm. Cloth

A close and up-to-date analysis of electronic devices is presented in this book, covering a broad field extending from the physical principles, through design and construction to practical uses of such devices and circuits based on them.

Semiconductor Devices. Measurements and Tests

G. GREEN

220 pp. 13 X 20 cm. Cloth

The book covers physical processes in semiconductor devices, their design and application. Describes the manufacture of semiconductor devices. Ample space is devoted to set-ups and techniques used to measure and test semiconductor devices. The book is designed as a text for vocational training.